High-end Video Streaming Qualityin the Wild : measuring and Predicting Satisfied User Ratio

Jingwen Zhu

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
Jingwen ZHU

High-end Video Streaming Quality in the Wild : Measuring and Predicting Satisfied User Ratio

Thèse présentée et soutenue à Nantes, le 09/09/2024
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Frédéric DUFAUX	Directeur de recherche CNRS, Université Paris-Saclay, France
Lu ZHANG	Maitre de conférences, HDR, INSA Renne, France

Composition du Jury :

Président :	Thomas MAUGEY	Directeur de recherche, Inria, France
Examinateurs :	Dietmar SAUPE	Professeur, Universität Konstanz, Allemagne
	Frédéric DUFAUX	Directeur de recherche CNRS, Université Paris-Saclay, France
	Lu ZHANG	Maitre de conférences, HDR, INSA Renne, France
Dir. de thèse :	Patrick LE CALLET	Professeur, Nantes Université, France

Abstract

The human eye cannot perceive small pixel changes in images or videos until a certain threshold of distortion. In the context of video compression, Just Noticeable Difference (JND) is the smallest distortion level from which the human eye can perceive the difference between reference video and the distorted/compressed one. Satisfied-User-Ratio (SUR) curve is the (complementary) cumulative distribution function of the individual JNDs of a viewer group. JND and SUR have been widely investigated for compressed image and video to use the least resources (e.g., storage and bandwidth) without damaging the Quality of Experience (QoE)

However, due to the extremely time-consuming nature of the JND subjective test, current available Video-Wise JND datasets are still very limited in terms of codec types, content resolution and dynamic range. In this thesis, we proposed a new AtHome protocol for subjective study on high-end video quality. We named this AtHome protocol "In-The-Wild" subjective test because it takes place in the diverse environment of participants' homes. We benchmark existing JND search methods and proposed a pre-processing step that significantly reduces the time required for JND searches. Using this optimized methodology, we collect new JND datasets for HEVC in HD-SDR and UHD-HDR videos.

After collecting the datasets, we introduce a new subjective data screening method named ZREC, which is simpler yet more effective than state-of-the-art methods, enhancing the reliability of the collected data. We propose methods for estimating confidence intervals for SUR, an often neglected but crucial aspect of SUR analysis. Additionally, we conduct a longitudinal study, a unique feature of our AtHome subjective test protocol.

By benchmarking widely used Video Quality Metrics (VQMs) on the JND datasets, we reveal their high content dependency at a given SUR threshold. We then propose a pipeline to predict SUR using VQMs as proxies and develop parameter-driven models to predict SUR using encoding parameters as proxies. Our proposed model is demonstrated to be less complex and more practical for streaming services.

Finally, we demonstrate the application of JND and SUR in optimizing streaming services. We analyze the bitrate costs for different SUR thresholds, showing that increasing SUR leads to an exponential increase in bitrate. Furthermore, we showcase how integrating

JND and SUR into bitrate ladder optimization can significantly save both bitrate and storage, ultimately enhancing the QoE for end-users.

Résumé

L'œil humain ne peut percevoir de petits changements de pixels dans les images ou les vidéos qu'à partir d'un certain seuil de distorsion. Dans le contexte de la compression vidéo, la Just Noticeable Difference (JND) est le plus petit niveau de distorsion à partir duquel l'œil humain peut percevoir la différence entre la vidéo de référence et la vidéo dégradée/compressée. La courbe du Satisfied-User-Ratio (SUR) est la fonction de distribution cumulative (complémentaire) des JND individuels d'un groupe de spectateurs. La JND et le SUR ont été largement étudiés pour les images et vidéos compressées afin d'utiliser le moins de ressources (par exemple, stockage et bande passante) sans nuire à la qualité d'expérience (QoE).

Cependant, en raison du temps considérable nécessaire pour les tests subjectifs de JND, les ensembles de données disponibles pour les vidéos JND sont encore très limités en termes de types de codecs, de résolution de contenu et de gamme dynamique. Dans cette thèse, nous avons proposé un nouveau protocole AtHome pour les études subjectives sur la qualité vidéo haut de gamme. Nous avons nommé ce protocole AtHome "In-the-Wild" car il se déroule dans l'environnement diversifié des domiciles des participants. Nous avons évalué les méthodes de recherche JND existantes et proposé une étape de prétraitement qui réduit considérablement le temps nécessaire pour les recherches JND. En utilisant cette méthodologie optimisée, nous avons collecté de nouveaux ensembles de données JND pour le HEVC en vidéos HD-SDR et UHD-HDR.

Après avoir collecté les ensembles de données, nous avons introduit une nouvelle méthode de filtrage des données subjectives nommée ZREC, qui est plus simple mais plus efficace que les méthodes actuelles, améliorant ainsi la fiabilité des données collectées. Nous proposons des méthodes pour estimer les intervalles de confiance pour le SUR, un aspect souvent négligé mais crucial de l'analyse SUR. De plus, nous menons une étude longitudinale, une caractéristique unique de notre protocole de test subjectif AtHome.

En évaluant les métriques de qualité vidéo (VQM) largement utilisées sur les ensembles de données JND, nous révélons leur forte dépendance au contenu à un seuil de SUR donné. Nous proposons ensuite un pipeline pour prédire le SUR en utilisant les VQM comme proxys et développons des modèles paramétriques pour prédire le SUR en utilisant les

paramètres d'encodage comme proxys. Notre modèle proposé s'avère moins complexe et plus pratique pour les services de streaming.

Enfin, nous démontrons l'application de la JND et du SUR dans l'optimisation des services de streaming. Nous analysons les coûts en débit pour différents seuils de SUR, montrant qu'une augmentation du SUR entraîne une augmentation exponentielle du débit. De plus, nous montrons comment l'intégration de la JND et du SUR dans l'optimisation de la bitrate ladder peut permettre de réaliser des économies significatives en termes de débit et de stockage, améliorant ainsi la QoE pour les utilisateurs.

Acknowledgement

I would like to first thank the two reviewers of my thesis: **Dr. Lu Zhang** and **Dr. Frédéric Dufaux**. I am grateful for their time and effort in reviewing my work. I would also like to thank **Dr. Thomas Maugey** and **Prof. Dr. Dietmar Saupe** for accepting to be jury members for my defense.

I would like to thank my supervisor **Prof. Dr. Patrick Le Callet** for his guidance and support throughout my Ph.D. journey. His expertise, vision, and commitment to rigorous and critical research have greatly influenced me. He encouraged me to push my limits, and I am grateful for the opportunity to work on this project under his continuous support and encouragement.

I would like to thank the people in Capacites SAS for their technical support, including **Dr. Suiyi Ling**, **Dr. Yoann Baveye**, **Dr. Pierre David**, **Charles Dormeval**, and **Dr. Anne-Flore Perrin**.

I would like to thank my collaborators inside and outside the lab, including **Dr. Ali Ak**, **Dr. Hadi Amirpour**, **Dr. Vignesh V Menon**, **Dr. Raimund Schatz**, and **Prof. Dr. Christian Timmerer**, who provided me with many valuable suggestions.

I would like to thank the members of the research group Image Perception and Interaction (**IPI**). Spending time with them was a pleasure, and they have been like a family to me, supporting me in both my research and my life. The IPI spirit will always be with me.

I would like to thank my Individual Monitoring Committee, including **Dr. Aladine Chetouani** and **Dr. Gérard Ramstein**, for following the progress of my thesis and providing valuable suggestions.

I would like to thank Amazon Prime Video for sponsoring this thesis and the collaborators from Prime Video, including **Dr. Sriram Sethuraman** and **Dr. Kumar Rahul**, for their support and suggestions, especially from an industry perspective.

Finally, I would like to thank my parents and Kevin for their companionship, trust, and encouragement during my Ph.D. journey.

Table of Contents

Abstract iii

Résumé v

Acknowledgement vii

Table of Contents ix

1 Introduction 1

1.1 Context and motivation . . . 1
1.2 Challenges and Research questions . . . 3
1.3 Organization and outlines . . . 6
1.4 Peer-reviewed publications . . . 8

2 In-the-Wild Subjective Testing for High-End Quality Assessment 11

2.1 Introduction . . . 11
2.2 AtHome subjective test system . . . 12
2.2.1 Displays Specifications . . . 13
2.2.2 Firestick and application . . . 14
2.3 Pre-qualification test . . . 15
2.3.1 Observers information . . . 16
2.3.2 Test sequences . . . 16
2.3.3 Test methodology . . . 17
2.4 Test environment's impact on subjective results . . . 18
2.4.1 MOS and CI . . . 19
2.4.2 Significant difference test . . . 21
2.4.3 Eliminated-by-Aspects . . . 22
2.5 Display impact on subjective results . . . 27
2.6 Summary . . . 28

3 "In-the-Wild" Subjective Study of VW-JND **31**
3.1 Introduction . . . 31
3.2 Background and motivation . . . 32
3.2.1 Subjective test for video quality assessment . . . 33
3.2.2 Satisfied User Ratio (SUR) of JND . . . 34
3.2.3 State of the art JND based datasets . . . 37
3.2.4 Other relevant works . . . 40
3.2.5 Motivation . . . 40
3.3 JND search methodology . . . 42
3.3.1 Related works . . . 42
3.3.2 Simulation and comparison of JND search methods . . . 44
3.3.3 Pre-processing of JCP . . . 50
3.4 Content selection . . . 52
3.5 Summary . . . 54

4 Subjective Data Analysis **55**
4.1 Introduction . . . 55
4.2 ZREC: robust recovery of mean and percentile opinion scores . . . 56
4.2.1 Background and motivation . . . 56
4.2.2 QoE Datasets . . . 59
4.2.3 Proposed Model . . . 59
4.2.4 Experiment Results . . . 63
4.2.5 Conclusion . . . 65
4.3 Uncertainty analyses of SUR . . . 66
4.3.1 Motivation . . . 66
4.3.2 Uncertainty estimation of $p\%\text{SUR}_{\text{emp}}$ . . . 67
4.3.3 Uncertainty estimation of SUR curve . . . 71
4.4 Longitudinal study of Subjective Data . . . 75
4.4.1 Test campaign management . . . 75
4.4.2 Observer behavior analysis . . . 77
4.5 Summary . . . 81

5 Objective Study of SUR **83**
5.1 Introduction . . . 83
5.2 Resolving power of VQM towards SUR . . . 84

5.2.1 VQM . . . 85
5.2.2 Experimental resuls . . . 86
5.3 Prediction of SUR using VQMs as proxy . . . 90
5.3.1 State-of-the-art methods . . . 93
5.3.2 $\Delta VMAF_{SUR(75\%)}$ prediction pipeline using VQMs . . . 95
5.3.3 Experimental Results . . . 97
5.3.4 Discussion . . . 99
5.4 Prediction of SUR using encoding parameters as proxy . . . 100
5.4.1 Parameter-driven model . . . 101
5.4.2 Further improvement of the prediction . . . 107
5.5 Summary . . . 114

6 Application of SUR: Streaming Optimization 117
6.1 Introduction . . . 117
6.2 Bitrate costs for enhanced user satifaction . . . 120
6.2.1 Expanding JND Datasets to other Codecs . . . 123
6.2.2 Bitrate as a Function of Satisfied User Ratio (SUR) . . . 126
6.2.3 Summary . . . 128
6.3 JND aware per-title bitrate ladder optimization . . . 129
6.3.1 `JASLA` Architecture . . . 130
6.3.2 Evaluation and Results . . . 133
6.3.3 Conclusions . . . 136
6.4 Summary . . . 136

7 Conclusion 139
7.1 Summary of contributions . . . 139
7.2 Limitations and Perspectives . . . 141

List of Abbreviations 143

Annexes 147
A System bias for SUR curve estimation? . . . 147
B Simple Staircase simulation . . . 149
C Quest+ Simulation . . . 150
D JND search methods accuracy benchmark . . . 150
E Intra campaign observer behavior analysis . . . 152

F Details of the Confidence Interval of the MLE 154
G Videos with higest and lowest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ 159
G.1 Videos with the 3 highest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ 159
G.2 Videos with the 3 lowest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ 160
H $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ prediction results with different seeds 160
I Spatial and Temporal Randomness 162
J Contributions to the Scientific Community 165

Bibliography **167**

Chapter 1

Introduction

Overview

Contents

1.1 Context and motivation 1
1.2 Challenges and Research questions 3
1.3 Organization and outlines 6
1.4 Peer-reviewed publications 8

1.1 Context and motivation

Human Visual System (HVS) cannot perceive small distortions. Just Noticeable Difference (JND) threshold is the minimum amount by which stimulus's intensity must be changed to produce a noticeable variation for HVS. Nowadays, with the increasing multimedia demand such as video streaming, JND plays an important role to reduce the resources (*e.g*, bandwidth, storage) consumption without decreasing the Quality of Experience (QoE) for end-users. In addition, JND has been widely employed in many other vision applications, including digital image/video processing, visual signal restoration/enhancement, and watermarking.

JND depends on 3 factors: (1) viewing conditions; (2) subjects; (3) video contents. This thesis investigates each of these factors:

1. **Viewing conditions:** The viewing conditions significantly impacts subjective test results for video quality experiments. Factors such as display type, ambient light, viewing distance and the enviroment [107] play crucial roles. Most subjective tests are conducted in controlled lab environments with well-calibrated displays and controlled ambient light and viewing distances, following International Telecommunication Union (ITU) recommendations [56, 58]. However, in real-life scenarios, people

do not watch videos in such controlled environments. To increase the ecological validity of subjective tests, we designed a subjective test system that allows participants to conduct tests at home with provided equipment and instructions. We named this system "In-the-Wild" subjective testing because it is conducted in diverse home environments with different displays. This system offers the opportunity to analyze the impact of the viewing conditions on the test results.

2. **Subjects:**

   The participants who conduct the subjective tests vary in their sensitivity to quality distortions. Factors such as age, educational background, visual acuity, and fatigue during the test affect the results. Analyzing the behavior of the subjects and cleaning the raw subjective data is crucial before drawing any conclusions. We propose a novel method to analyze subjects' behavior during the tests and clean the raw data. Conducting the JND subjective test on a group of subjects (*e.g.,* 30 participants) for a given video content can provide the Satisfied User Ratio (**SUR**), which is the percentage of subjects who cannot perceive the distortion compared to the anchor. The concept of SUR was proposed in [138] to account for differences in subjects' JND thresholds. This thesis extends this concept to a more general form for different JND proxies such as QP, bitrate, and VMAF.

3. **Video contents:**

   Video contents differ in terms of resolution, dynamic range, and intrinsic characteristics such as motion, texture, and color. For the same viewing environment and subjects, different video contents will have different JND thresholds. This is because various content types react differently to compression due to spatial and temporal complexities, and some content can mask the perception of distortion by the HVS, known as the masking effect [50]. After collecting and cleaning the subjective test data, we analyze the impact of video content on the SUR of JND and develop prediction models for SUR and JND with different proxies. This thesis focuses on Video-on-Demand content and high-end video quality.

The first step in modeling JND is to collect subjective datasets. Although several JND datasets are available in the literature, the JND datasets for videos are still limited in terms of resolution [63], dynamic range [138] and other factors. Additionally, these datasets are typically collected in well-controlled lab environments, which may not reflect real-life scenarios. Therefore, it is essential to expand the collection of JND datasets for videos to address these limitations.

To identify the JND video in a series of videos with varying distortion levels for a given anchor/reference, these videos need to be compared with the anchor. Wang *et al.* [138] proposed a binary search method to conduct the JND search which can help to reduce the number of comparison during the JND search. However, compared to standard subjective tests for image/video quality (such as Absolute Category Rating (ACR) or Degradation Category Rating (DCR) [61]), subjective JND testing is significantly more time-consuming because it requires tracking the HVS threshold. Given that subjective tests are time and resource-intensive, developing JND prediction models is essential.

Finally, integrating the JND prediction model into video streaming can significantly reduce resource consumption without compromising the QoE for end-users. By selecting the appropriate JND threshold for streaming, we can optimize bitrate and storage costs effectively. It is crucial to investigate the impact of different SUR thresholds on streaming bitrate and quantify the potential savings in bitrate and storage costs through JND-aware bitrate ladder optimization.

1.2 Challenges and Research questions

In this thesis, we investigate both the subjective and objective modeling of JND by considering these 3 factors. We also analyze the potential impact of resource savings by integrating JND into video streaming systems. We formulated the following research questions:

1. **RQ1: How to Design a More Ecologically Valid Subjective Test System for High-End Video Quality?**

   Traditional subjective tests for high-end video quality are usually conducted in well-controlled laboratory environments. These environments include well-calibrated displays, fixed viewing distances, and controlled ambient lighting. However, in real-life scenarios, people do not watch videos under such controlled conditions.

   Crowdsourcing platforms like Amazon Mechanical Turk (AMT) and Prolific have been widely used to conduct subjective tests for video quality [48, 2, 106], providing conditions that are closer to real-life scenarios. Nevertheless, these platforms have limitations, particularly regarding display choice. For example, testing video quality on a 55-inch TV is challenging on these browser-based platforms.

   Therefore, the research question arises: How can we design a more ecologically valid subjective test system for high-end video quality? This question seeks to develop

methods and systems that better replicate real-life viewing conditions while maintaining the rigor and reliability of traditional lab-based tests.

2. **RQ2: What is the Impact of the Viewing Conditions on the Subjective Test Results?**
   We collect subjective test data from participants at home using our provided equipment and instructions. This research question aims to understand the impact of the viewing conditions on the results, including differences between "in lab" and "at home" settings, and how the type of display affects the outcomes. By comparing data from controlled lab environments and diverse home settings, we seek to identify and analyze variations in test results attributable to these different conditions.
3. **RQ3: What is the Impact of Conducting Subjective Tests Over the Long Term on Test Results?**
   Since we collect subjective test data at each participant's home, the tests can be conducted over an extended period. Participants can complete the tests at their convenience, with a daily time limit to avoid fatigue, and are required to participate for a specified number of days (e.g., 9 days with 45 minutes per day) for each test campaign. This unique design allows us to perform a longitudinal study on the impact of long-term subjective testing on the results, providing insights into how extended testing durations affect participant responses and overall test outcomes.
4. **RQ4: Which Subjective Test Methodology is Best for Tracking the JND Threshold for Video Quality?**
   The binary search method is the most widely used subjective test methodology for JND dataset collection [138]. However, there are other psychophysical methods, such as the method of limits [35], the staircase method [15], and Quest+ [142], which can be used to determine JND. Determining the most suitable methodology for tracking the JND threshold for video quality remains an open question. By benchmarking these different methods, we aim to identify the most effective approach. This will allow us to further optimize subjective test methodologies, given that collecting JND datasets is a very time-consuming process.
5. **RQ5: How to Analyze Subjects' Behavior from the Subjective Data and Clean the Raw Subjective Data?**
   After collecting the raw subjective data, it is essential to analyze the subjects' behavior, such as bias and inconsistency, and clean the data. This research question aims to develop a novel method for analyzing subjects' behavior and cleaning the

raw subjective data. This method will help identify and remove outliers, ensuring the reliability and validity of the subjective test results.

6. **RQ6: What is the Impact of Subjective Data Screening on Learning-Based Prediction Models?**
   After cleaning the raw subjective data, the cleaned data can be utilized to train a learning-based prediction model for JND and SUR. This research question aims to explore the effect of subjective data screening on the performance of such learning-based prediction models. By comparing the results of models trained with and without cleaned data, we can assess the effectiveness of the data screening process and its influence on model performance.
7. **RQ7: How to Estimate the Uncertainty of SUR Obtained from Subjective Tests?**
   We already know that relying solely on the Mean Opinion Score (MOS) is not sufficient, as it overlooks the diversity in subjective ratings [47]. Consequently, the uncertainty of the MOS has been widely investigated in the literature [119, 28, 25]. Similarly, the uncertainty of SUR is equally crucial yet remains understudied. In this research question, our aim is to estimate the uncertainty of SUR obtained from the subjective test. This will provide us a more comprehensive understanding of the subjective test results.
8. **RQ8: How Effectively Can Current Widely Used Video Quality Metrics (VQMs) Reflect SUR?**
   To what extent can the current widely used Video Quality Metrics (VQMs) reflect the SUR? Using VMAF as a case study, can we identify a threshold VMAF value at which 75% of observers cannot perceive a quality difference compared to the pristine video? We refer to this as the resolving power of VQM towards SUR. Which VQM has the best resolving power towards SUR?
9. **RQ9: How to Develop a Prediction Model for SUR and JND without Extensive Recompression?**
   This research question addresses the challenge encountered in previous JND/SUR prediction models, where a large number of Processed Video Sequences (PVSs) are required as input, proving time and storage-intensive for video streaming service providers. We aim to explore a method that relies solely on Source Video Content (SRC) as input, eliminating the need for re-compression, to predict SUR curves with a given codec. This question is based on the assumption that SUR curves are

predominantly influenced by content features, such as the masking effect of spatial and temporal randomness [137].

10. **RQ10: What is the impact on bitrate when selecting different SUR thresholds for streaming?**
    How does the bitrate change when selecting different SUR thresholds? Specifically, how much additional bitrate is required to increase the SUR from 75% to 95%? This research question aims to provide insights into the trade-off between bitrate and SUR, enabling service providers to selecte the appropriate SUR threshold for streaming.
11. **RQ11: How Much Bitrate and Storage Cost Can Be Saved with JND-Aware Bitrate Ladder Optimization?**
    Bitrate ladders are utilized in adaptive streaming to offer varying quality levels for different network conditions. Yet, these ladders often contain redundant quality levels. This research question investigates the potential savings in bitrate and storage costs achievable through JND-aware bitrate ladder optimization.

1.3 Organization and outlines

The organization of the thesis and the corresponding research questions of JND are shown in Figure 1.1. In **Chapter 1**, we introduce the context and motivation of the thesis, outline the challenges and research questions, and provide an overview of the thesis structure and peer-reviewed publications.

In **Chapter 2**, we address **RQ1** by introducing our "AtHome" subjective test system. We conducted pre-qualification tests to ensure the reliability of our "AtHome" test settings and analyze the impact of the test environment on the subjective test results (**RQ2**).

In **Chapter 3**, we compare the state-of-the-art JND datasets and address **RQ4** by benchmarking different JND search methodologies through simulations. We also describe how we optimized the JND search methodology and collected our JND datasets using the AtHome subjective test system.

Chapter 4 addresses **RQ5**, **RQ6**, and **RQ7**. We propose a novel method that is simpler yet more efficient than the state-of-the-art methods for analyzing subjects' behavior and cleaning raw subjective data. Additionally, we investigate the impact of subjective data screening on learning-based prediction models and estimate the uncertainty of the SUR obtained from subjective tests.

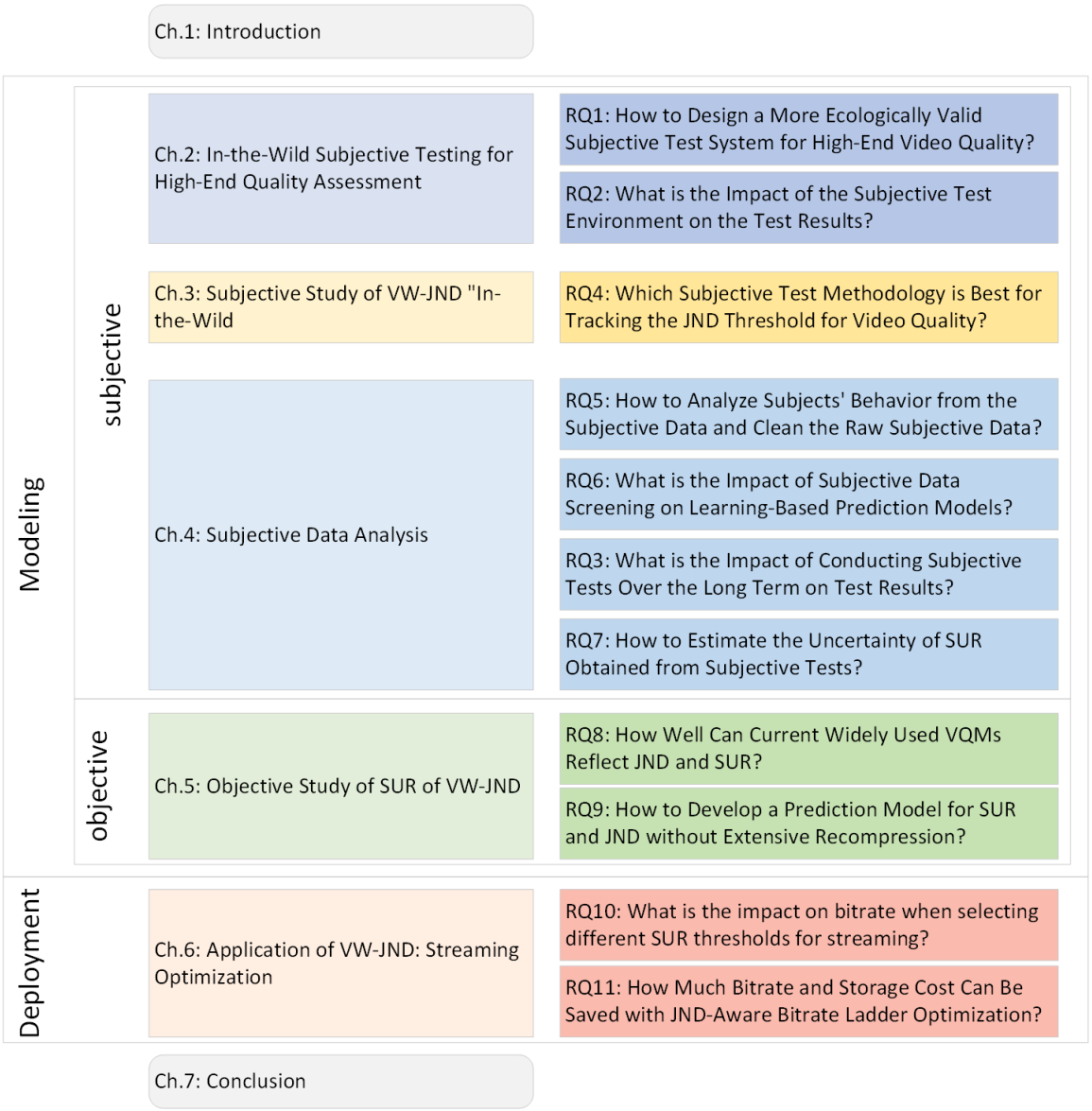


Figure 1.1 – Organization of the thesis and the corresponding research questions

In **Chapter 5**, we first address **RQ8** by analyzing how the currently widely used VQMs reflect JND and SUR. We then address **RQ9** by developing a prediction model for SUR and JND using different proxies (e.g., VQM, encoding parameters), avoiding the resource-intensive and impractical process of massive re-compression for the industry.

Chapter 6 addresses **RQ10** and **RQ11** by investigating the impact of selecting different SUR thresholds for streaming on bitrate, and how much bitrate and storage cost can be saved with JND-aware bitrate ladder optimization.

Chapter 7 concludes the thesis by summarizing the contributions and limitations, and provides directions for future research.

1.4 Peer-reviewed publications

Publications presented in the thesis:

1. **Jingwen Zhu**, Patrick Le Callet, Anne-Flore Perrin, Sriram Sethuraman, and Kumar Rahul. "On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video." In 2022 IEEE International Conference on Image Processing (ICIP), pp. 4213-4217. IEEE, 2022.
2. **Jingwen Zhu**, Anne-Flore Perrin, and Patrick Le Callet. "Subjective test methodology optimization and prediction framework for Just Noticeable Difference and Satisfied User Ratio for compressed HD video." In 2022 Picture Coding Symposium (PCS), pp. 313-317. IEEE, 2022.
3. **Jingwen Zhu**, Ali Ak, Charles Dormeval, Patrick Le Callet, Kumar Rahul, and Sriram Sethuraman. "Subjective Test Environments: A Multifaceted Examination of Their Impact on Test Results." In Proceedings of the 2023 ACM International Conference on Interactive Media Experiences (IMX), pp. 298-302. 2023.
4. Menon, Vignesh V., **Jingwen Zhu**, Prajit T. Rajendran, Hadi Amirpour, Patrick Le Callet, and Christian Timmerer. "Just noticeable difference-aware per-scene bitrate-laddering for adaptive video streaming." In 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1673-1678. IEEE, 2023.
5. **Jingwen Zhu**, Ali Ak, Patrick Le Callet, Sriram Sethuraman, and Kumar Rahul. "ZREC: robust recovery of mean and percentile opinion scores." In 2023 IEEE International Conference on Image Processing (ICIP), pp. 2630-2634. IEEE, 2023.
6. **Jingwen Zhu**, Hadi Amirpour, Raimund Schatz, Christian Timmerer, and Patrick Le Callet. "Enhancing Satisfied User Ratio (SUR) Prediction for VMAF Proxy through Video Quality Metrics." In 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), pp. 1-5. IEEE, 2023.
7. Menon, Vignesh V., **Jingwen Zhu**, Prajit T. Rajendran, Samira Afzal, Klaus Schoeffmann, Patrick Le Callet, and Christian Timmerer. "Optimal Quality and Efficiency in Adaptive Live Streaming with JND-Aware Low latency Encoding." In Proceedings of the 3rd Mile-High Video Conference (MHV), pp. 61-67. 2024.

8. Amirpour, Hadi, **Jingwen Zhu**, Raimund Schatz, Patrick Le Callet, and Christian Timmerer. "Exploring Bitrate Costs for Enhanced User Satisfaction: A Just Noticeable Difference (JND) Perspective." In Data Compression Conference. 2024.
9. **Jingwen Zhu**, Hadi Amirpour, Raimund Schatz, Christian Timmerer, and Patrick Le Callet. "Beyond Curves and Thresholds - Introducing Uncertainty Estimation to Satisfied User Ratios for Compressed Video." In 2022 Picture Coding Symposium (PCS). IEEE, 2024. (*Best Paper Award*)

Other publications:

1. **Jingwen Zhu**, and Patrick Le Callet. "Just noticeable difference (JND) and satisfied user ratio (SUR) prediction for compressed video: research proposal." In Proceedings of the 13th ACM Multimedia Systems Conference (MMSys), pp. 393-397. 2022.
2. **Jingwen Zhu**, Suiyi Ling, Yoann Baveye, and Patrick Le Callet. "A framework to map vmaf with the probability of just noticeable difference between video encoding recipes." In 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1-5. IEEE, 2022.
3. **Jingwen Zhu**, Hadi Amirpour, Vignesh V. Menon, Raimund Schatz, and Patrick Le Callet. "Elevating Your Streaming Experience with Just Noticeable Difference (JND)-based Encoding." In Proceedings of the 2nd Mile-High Video Conference (MHV), pp. 128-129. 2023.
4. Liu, Jiawen, **Jingwen Zhu**, and Patrick Le Callet. "Bridge the Gap between Visual Difference Prediction Model and Just Noticeable Difference Subjective Datasets." In 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), pp. 1-5. IEEE, 2023.
5. Leszczuk, Mikołaj, Lucjan Janowski, Jakub Nawała, **Jingwen Zhu**, Yuding Wang, and Atanas Boev. "Objective Video Quality Assessment and Ground Truth Coordinates for Automatic License Plate Recognition." Electronics 12, no. 23 (2023): 4721.
6. Amirpour, Hadi, **Jingwen Zhu**, Patrick Le Callet, and Christian Timmerer. "A Real-Time Video Quality Metric for HTTP Adaptive Streaming." In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3810-3814. IEEE, 2024.

Chapter 2

In-the-Wild Subjective Testing for High-End Quality Assessment

Overview

Contents

2.1 Introduction 11
2.2 AtHome subjective test system 12
2.2.1 Displays Specifications 13
2.2.2 Firestick and application 14
2.3 Pre-qualification test 15
2.3.1 Observers information 16
2.3.2 Test sequences 16
2.3.3 Test methodology 17
2.4 Test environment's impact on subjective results 18
2.4.1 MOS and CI 19
2.4.2 Significant difference test 21
2.4.3 Eliminated-by-Aspects 22
2.5 Display impact on subjective results 27
2.6 Summary 28

Part of this chapter has been published in research papers [160].

2.1 Introduction

Traditional subjective tests for high-end quality are usually conducted in well-controlled laboratory environments, which include a well-calibrated display, a fixed viewing distance, and controlled ambient light, among other factors. However, the recent pandemic at the

(a) InLab (b) AtHome

Figure 2.1 – InLab and AtHome subjective test environment

beginning of this thesis made it difficult to conduct such tests in a laboratory. Consequently, we decided to conduct the subjective tests at the homes of the participants.

The advantage of doing the test at home is that we can collect the data in a more ecological valid way. The participants are watching the video in their own environment, which is more similar to the real life scenario. Furthermore, we used different brands of TV to conduct the test, helping us to understand the impact of the display on the subjective test for high-end quality.

We refer to this type of subjective test as an **"In-the-wild"** subjective test because it takes place in the diverse environment of participants' homes, featuring a variety of displays, rather than in a well-controlled laboratory setting. We name the two types of subjective tests enviroment as "InLab" and "AtHome". Figure 2.1 shows the two different environments for the subjective tests.

We developed a complete system to conduct reliable subjective tests at the home of each participant (see Section 2.2). In Section 2.3, we described the pre-qualification test which is desinged to ensure the correct setting of "AtHome" enviroment. In Section 2.4, we analyzed the impact of the test environment (AtHome and InLab) on the subjective test results. Similarly, in Section 2.5, we analysed the impact of different displays on the subjective test results.

2.2 AtHome subjective test system

Due to the COVID-19 pandemic, we implemented a pipeline to conduct the subjective tests at participants' homes. This strategy enabled us to gather data in a setting more ecologically valid, with participants viewing videos in real-life scenarios. Additionally, we employed a variety of TV brands in the tests, providing insights into how the display impacts the subjective assessment of high-end quality.

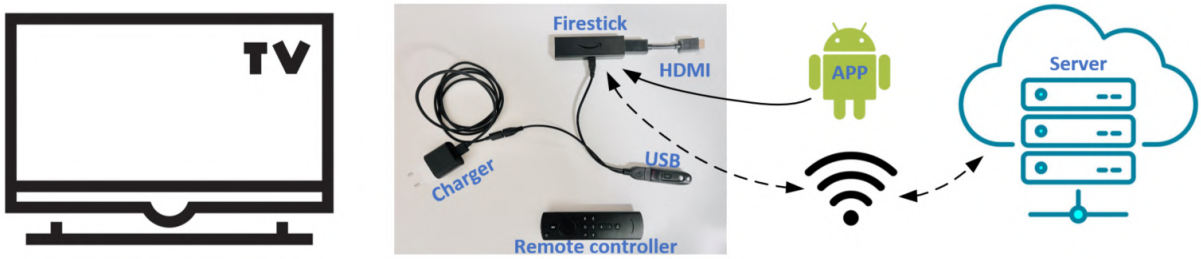


Figure 2.2 – A simplified AtHome subjective test system

Figure 2.2 outlines the simplified system for the AtHome subjective test. Participants received the assigned TV, firestick, and other necessary materials. They set up the TV, connected the provided firestick to the TV via HDMI, and ensured the firestick had an internet connection through Wi-Fi. The application was pre-installed on the firestick by our team. Participants then logged in with their accounts on the application and initiated the test. Subsequently, the application sent ratings and playback logs to the server, from which we extracted data for further analysis. The detailed system is described below.

2.2.1 Displays Specifications

The TVs used in our AtHome subjective test are detailed in Table 2.1. The selection

Table 2.1 – TV Models detials for the AtHome subjective test

Brand	Reference	No. of TVs	Size	Resolution	Screen Type	Screen Backlight
SONY	KD-55XH8096	4	55 inches	3840 x 2160	LCD	LED
LG	Nanocell 55NAN091	2	55 inches	3840 x 2160	LCD	LED
SAMSUNG	UE55TU8075U	2	55 inches	3840 x 2160	LCD	LED
SAMSUNG	QE55Q74TATXXC	2	55 inches	3840 x 2160	LCD	QLED

of these displays was influenced by both the market share of different brands and the project's budget. All the TV selected are LCD display with LED backlight. Participants were instructed to set up the displays in their living rooms for comfortable video viewing. Each home had two participants, both of whom had passed a pre-experiment vision check to ensure they possessed normal or corrected-to-normal visual acuity.

We configured all the displays before delivering to the participants, ensuring standard settings and removing post-processing features like denoising and motion flow. While participants were allowed to use the TV outside of the test for personal purposes, they were requested to keep the settings unchanged. We measured the displays using an X-Rite

i1Display Pro colorimeter and can be used to analyse the impact of the display on the subjective test results.

Participants were instructed to position themselves in the middle, directly in front of the TV, maintaining a distance of 2 meters for optimal viewing of HD content. This distance was approximately three times the height of the screen. For UHD content, the recommended viewing distance was 1.5 times the height of the screen, which equates to approximately 1 meter from the TV.

Ambient light, especially light directly hitting the screen, can significantly affect luminance, particularly for black levels due to display reflectivity. The display contrast decreases with an increase in ambient light, and this effect is also influenced by the anti-reflective coating of the display and room lighting geometry [71]. Uncontrolled ambient light in crowdsourcing subjective tests increases the variance of the Mean Opinion Scores (MOS). To mitigate the impact of ambient light, participants are instructed to close curtains or blinds and turn off lights facing the screens. Additionally, participants may turn on some lights in the room to create a dimly lit atmosphere while avoiding complete darkness.

2.2.2 Firestick and application

The Amazon Fire TV Stick, a portable device that plugs into a TV's HDMI port, provides access to streaming services like Amazon Prime Video, Netflix, and Disney+. It also supports gaming and apps such as YouTube and Spotify, connecting to the internet via Wi-Fi and powered by a USB cable linked to the TV or a wall outlet as shown in the Figure 2.2.

The advantage of using the Fire TV Stick is that it provides a consistent platform for video playback across different TV brands, eliminating the need to navigate different operating systems on various smart TVs. In our experiment, we used the Fire TV Stick 4K Max, which runs on Fire OS 8, based on Android 10 (API level 29).

We configured the Fire TV Stick to automatically upscale the video to match the resolution of the source content (i.e., viewing resolution). The settings of the Fire TV Stick can significantly impact the visualization of the videos; we list the important settings in Table 2.2.

We developed an application for the subjective test using the Android native player. The application is designed to be user-friendly and easy to navigate, providing a simple interface for participants to log in, initiate the test, and rate the videos.

Table 2.2 – Settings of the Fire TV Stick for the subjective test

Content Type	Resolution	Color Depth	Dynamic Range
SDR	Up to 4K Ultra HD	8 bit	Deactivate HDR
HDR10	Up to 4K Ultra HD	Up to 10 bit	Always HDR

In crowdsourcing subjective tests, reference videos are often compressed (e.g., CRF of 18 in [13], CRF of 12 in [63]) to reduce file size [40] and ensure smooth playback. However, with the Fire TV Stick 4K Max, we can play videos at a much lower CRF (CRF of 1 with the Prime Video in-house encoder) and a bitrate exceeding 400Mbps, without filesize constraints. This allows the reference videos to closely resemble the original quality, aligning with our focus on high-end quality.

The application is designed to handle various test types, including Degradation Category Rating (DCR), Absolute Category Rating (ACR), and more. The test type remains transparent to the participants, with instructions provided each time they log in. We have the flexibility to change the test type for each participant on the server side, and the application will seamlessly adjust to the assigned test type. The application sends both ratings and playback logs to the server, from which we extract data for further analysis. Our team pre-installed the application on the Fire TV Stick, and participants were instructed to log in with their accounts and initiate the test.

As shown in Figure 2.2, a USB stick is connected to the Fire TV Stick. All the videos to be evaluated are stored on the USB stick and are fully loaded before playback. This approach eliminates the need for participants to download the videos, which could potentially lead to issues such as network congestion and slow download speeds.

2.3 Pre-qualification test

The goal of the pre-qualification test is to ensure participants can successfully conduct the test in a qualified environment at home. We conducted the same test in both a well-controlled laboratory environment with a well-calibrated display and at participants' homes with various displays (refer to Section 2.2.1). By comparing the results of these two tests, we confirm participants' ability to conduct the test in a qualified setting. Additionally, this comparison allows us to analyze the impact of the test environment (AtHome *vs.* InLab) on the subjective test results.

2.3.1 Observers information

We recruited 20 participants for the AtHome subjective test. Each pair of participants shared the same household environment and display. These pairs typically consisted of family members or friends. The observers were naive, and the experiment duration was compensated by gift cards. They varied in age, nationality, and educational backgrounds, as shown in Table 2.3. Every participant passed the pre-experiment vision check to make sure that they have normal or corrected-to normal visual acuity.

Table 2.3 – Repartition of Ages and Genders in the Subjective Tests

Age				**Gender**	
Average	**Std**	**Min**	**Max**	**Male**	**Female**
36.3	12.4	20	63	9	11

In Section 2.2.1, we outlined instructions for participants to set up the displays, control the viewing distance, and manage ambient light. While participants were carefully selected from a pool of trusted panelists—who have not been flagged as spammers in previous subjective tests—we still needed to ensure adherence to instructions and maintain a qualified testing environment. Hence, we designed a pre-qualification test before conducting further subjective test.

2.3.2 Test sequences

Content selection plays a crucial role in enhancing the efficiency of subjective tests by enabling the selection of representative contents [120]. Spatial information (SI), Temporal information (TI) [55] and Ambiguity [88] are computed on 229 HD (1080p) videos contents provided by Amazon Prime Video. K-means clustering was applied to the extracted features to group the contents based on their similarities. From these content clusters, 10 HD contents C_i were selected, with each content having a duration of 10 seconds.

Each selected contents are compressed by High Efficiency Video Coding (HEVC) with 3 different encoding resolutions (1080p, 720p and 540p). For each encoding resolutions, 13 different level of distortions are used. There are in total 39 (3×13) encoding recipes $R_{i,j}$ for each content C_i. However, it is time and money consuming to conduct subjective test on all the generated Processed Video Sequence (PVS). To select significant PVS for the subsequent subjective test, we conducted a 2-direction JND search by experts (golden-eyes) as described in Figure 2.3, following the framework proposed in [155].

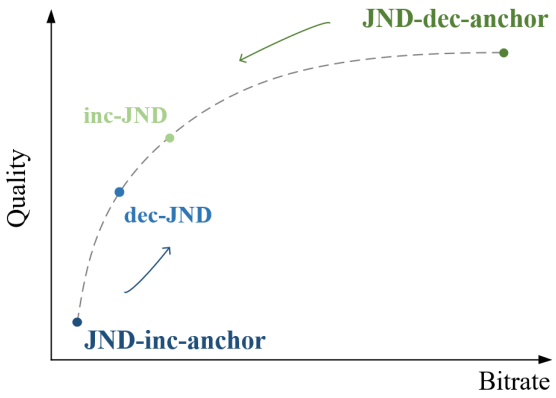


Figure 2.3 – 2-direction JND search

For each resolution, we selected the PVS with the highest quality as the anchor (referred to as JND-dec-anchor) and searched through the remaining PVS to identify the point at which the observer just begins to perceive a difference in quality between the anchor and the PVS (referred to as inc-JND points). Likewise, we selected the PVS with the lowest quality as the JND-inc-anchor to identify the JND points where the observer first perceives a quality difference. There are a total of 12 stimulus for each content, consisting of one source (SRC) and 11 PVS.

2.3.3 Test methodology

The Degradation Category Rating (DCR) test methodology are used for the pre-qualification test. The DCR test involved presenting the test sequences in pairs, where the first stimulus in each pair was always the source reference, and the second stimulus was the PVS of the same content. Five-level scale for rating was used as recommended in ITU-T P.910 [55] as shwon in Table 2.4.

Table 2.4 – Degradation Category Rating Scale (DCR)

Rating	Description
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The stimuli were presented to the subjects in a random order, with a constraint that the same content was not presented successively. The entire test took approximately 45 minutes to complete.

We conducted the same DCR test in both the InLab and AtHome environments (see Figure 2.1). Both share identical contents and experiment design. The only difference between the two experiments was the experiment setting where one of the experiments was conducted in a controlled laboratory environment and the other was in home environment of each participants.

The InLab experiement specifications are as follows:

— Display: a 4K calibrated UHD Grundig Finearts 55 FLX 9492 SL with a 55-inch screen size.

— Viewing distance: 3H for HD (1080p) video, with H the height of the screened video, as recommended in ITU-R BT.1769 [59]

— Ambient light: the illuminance level of the subjective environment was set as recommended by ITU-R BT.2013-1 [60]

— Subjects: A total of 24 participants, who were non-experts in subjective experiments, image processing, or related fields, took part in the study. All participants had either normal or corrected-to-normal visual acuity, which was ensured prior to the experiment using a Monoyer chart. Ishihara color plates were used to test color vision, and all viewers passed the pre-experiment vision check. There is no overlap between the participants in the InLab and AtHome experiments to avoid any potential bias.

Details regarding the AtHome experiment specifications can be found in 2.2.

2.4 Test environment's impact on subjective results

Previous studies [72, 77] have identified several factors that can significantly impact Quality of Experience (QoE) for multimedia content. These factors include, but are not limited to, video quality, device, observer's emotion *etc.*

The physical experiment environment is a major factor that influences QoE, and standardization efforts have been made to propose methodologies and recommendations for experiment conditions in subjective QoE studies. For instance, ITU has developed standards such as BT.500 [56], BT.910 [58], and BT.913 [57].

In a previous study by Jumisko-Pyykö *et al.* [65], it was found that the acceptability of content varied significantly when measured in laboratory environment compared to real-life scenarios. The results indicated that subjects were more critical during laboratory experiments when evaluating the acceptability of mobile videos. In another study, Li *et al.* [77] investigated the influence of devices on Quality of Experience (QoE), focusing on acceptability and annoyance in video streaming. Their analysis revealed a significant impact on measured QoE when comparing devices, such as Tablets vs. TVs. Additionally, Ak *et al.* [3] conducted subjective tests to assess the influence of context on video streaming QoE. The findings indicated that the remaining data in a mobile data plan context could impact participants' opinion scores.

In this section, we analyze the impact of the experiment environment (AtHome *vs.* InLab) on the results obtained from the pre-qualification test (see Section 2.3). We compare the Mean Opinion Scores (MOS) and Confidence Intervals (CI) of the MOS collected from both enviroments. Additionally, we perform an ANOVA test to assess the statistical impact of the experiment environment on the subjective test results. Finally, we conduct an advanced analysis, specifically Eliminated-by-Aspects (EBA) method to measure the influence as a function of MOS.

2.4.1 MOS and CI

We analyzed the correlation between the MOS collected from the two environments (InLab and AtHome) and compare their Confidence Intervals (CI). We expect a smaller CI for the InLab experiment. We use 3 different MOS recovery (observer screening) methodologies to analyze the correlation between the MOS collected in InLab and AtHome experiments. Summary of each method can be found below:

BT500: ITU-R BT.500 Recommendation [56] defines the simple and commonly used observer screening procedure. Subjects are rejected based on the number of opinion scores outside of the predefined amount of standard deviation range of the population. If a subject found to be an outlier, all of his/her opinions are removed from the dataset. MOS is calculated as the mean of remaining subjects.

P910: is specified in ITU-R P.913 Recommendation clause 12.6 [57] (also refferred to as **P913-12.6**) and defines a procedure where MOS is recovered by bias removal and subject inconsistency based weighted average. The procedure defines the individual opinion scores of a subject as the combination of subject bias, inconsistency and the true quality of the stimuli. The approach simultaneously addresses and resolves these three parameters.

ZREC: is proposed in [161]. It relies on estimating subject bias and inconsistency to recover the MOS. It doesn't require any solver and the evaluations show a smaller CI over the tested dataset compared to alternative methods. For further details and insights into MOS recovery methods, refer to Section 4.2.

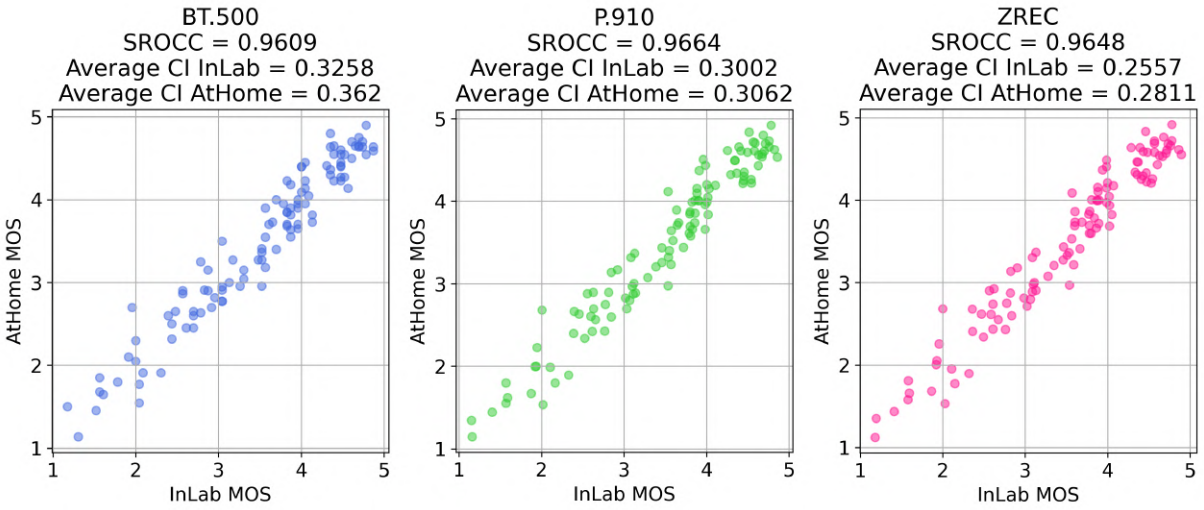


Figure 2.4 – Scatter plots between InLab MOS and AtHome MOS with 3 different MOS recovery (observer screening) methods. Below the title of each plot, Spearman's Rank Order Correlation Coefficients (SROCC) and average CI 95% for InLab and AtHome experiments are given.

With each method describe above, MOS from the InLab and AtHome experiments and their CI (95%) are calculated. Figure 2.4 presents the results as a scatter plot between the InLab and AtHome MOS for each method. SROCC values indicate that the MOS acquired from both enviroments are highly correlated with all MOS recovery methods. On another front, we observe slightly lower CI for the InLab MOS compared to AtHome MOS. Considering the uncontrolled experiment environments in AtHome experiment, the results are not surprising. With more sophisticated MOS recovery methods, we can acquire lower CIs for both experiments however the slightly higher CI for AtHome experiment remains true.

We also analysed the Standard Deviation of Opinion Scores (SOS) [47] for each stimulus. The SOS analyses is widely used to compare the annotation quality of different subjective test [71, 40, 147, 82]. SOS hypothesis suppose the variance of the opinion scores $sigma^2$ follows the Eq.(2.1).

$$SOS(MOS)^2 = a(-MOS^2 + 6 \times MOS - 5) \tag{2.1}$$

As shown in Figure 2.5, for the InLab test, we obtain $a = 0.2177$, a value consistent with other in-lab tests in video compression and streaming [47]. It is noteworthy that the a value for the AtHome test is higher than that for the InLab test, aligning with our earlier observation that the AtHome test exhibits a higher Confidence Interval (CI) than the InLab test.

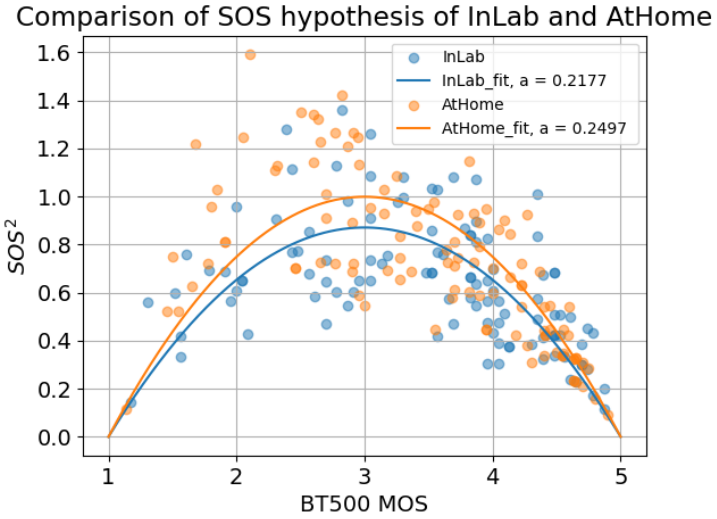


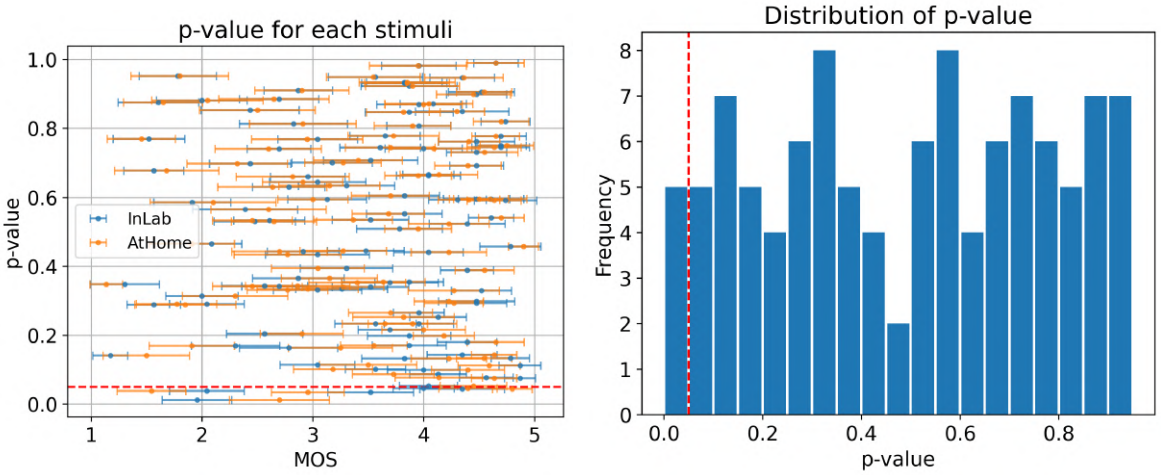
Figure 2.5 – Comparison of SOS hypothesis of InLab and AtHome experiments

2.4.2 Significant difference test

In the previous section, we analyzed the correlation between the MOS collected from the InLab and AtHome experiments and the SOS scores respectively. In this section, we perform a significant difference test to determine whether the opinion scores and the corresponding CI collected from the two environments are significantly different.

Significant difference test for Opinion Scores

For each stimulus, we performed a significant difference test for the opinion scores between the InLab and AtHome experiments. As discussed in Section 2.4.1, it can be observed that the variances of the opinion scores for the two environments are different. Therefore, we employed the Welch's t-test [125], which is an adaptation of Student's t-test and is more suitable when variances are not assumed to be equal. The results are depicted in Figure 2.6.



(a) Welch's t-test p-value for opinion scores

(b) Distribution of p-values

Figure 2.6 – Significant difference test for per stimulus' opinion scores

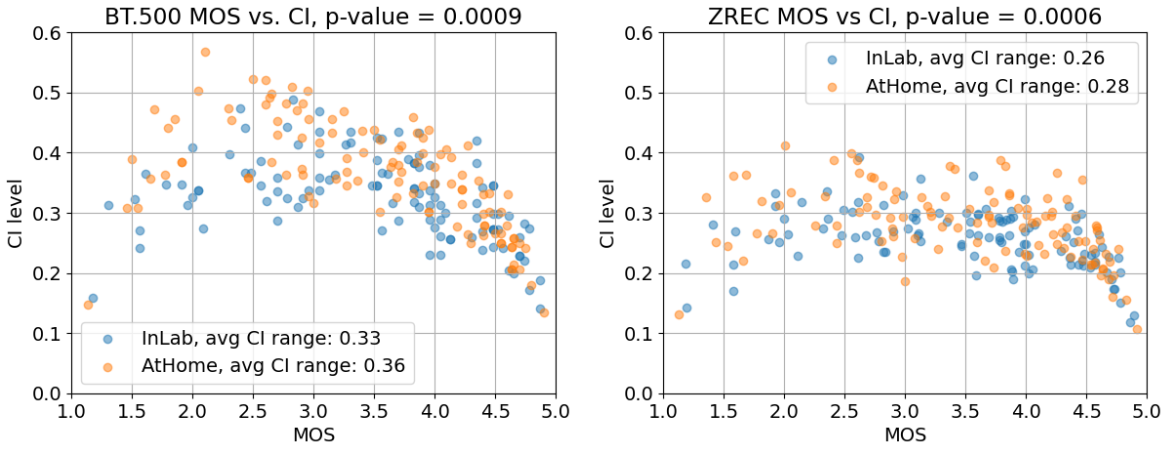
It can be observed that among the 110 stimuli, only 5 stimuli have p-values less than 0.05 (see Figure 2.6b), indicating that the opinion scores in the InLab and AtHome experiments are not significantly different for most cases. Among these 5 stimuli, 3 of them have AtHome MOS higher than InLab MOS, while the other 2 have InLab MOS higher than AtHome MOS (see Figure 2.6a). These results indicate that the experiment environment has a limited impact on the opinion scores.

Significant difference test for CI

Similar to the SOS analyses, Figure 2.7 shows the relationship between the Confidence Interval (CI) level and the Mean Opinion Score (MOS) for both the InLab and AtHome experiments. The significant difference test (Welch's t-test) indicate that the CI level is significantly different for the two environments (Figure 2.7a and 2.7b show the MOS obtained by BT500 [56] and ZREC [161] respectively). It is also observed that using ZREC [161] to recover the MOS can reduce the CI level for both the InLab and AtHome experiments. For more details about the opinion score recovery method, readers can refer to Section 4.2.

2.4.3 Eliminated-by-Aspects

In this part, we rely on Eliminated-by-Aspects (EBA) analysis the impact of experiment settings on the QoE and quantify this impact.



(a) CI vs. BT500 MOS

(b) CI vs. ZREC MOS

Figure 2.7 – CI vs. MOS for InLab and AtHome experiments

EBA model is proposed by Tversky [133]. It can be used to analyze the subgroups containing similar stimuli. It assumes that, in a pair comparison subjective experiment, a subject prefer a stimulus than another stimulus in comparison due to set of attributes being present in the higher quality stimulus.

In QoE experiments, each video sequence i has a quality attribute defined by q_i. If no other influences are considered, the MOS can be represented as $f_m(q_i)$, where f_m is typically a logarithmic mapping function [77]. In our case where experiment were repeated under home and laboratory settings, these settings also have their attributes defined as d_{Home} and d_{Lab}. Therefore, the measured QoE in each experiment can be represented as $f_m(q_i + d_i)$ where d_i is either d_{Home} or d_{Lab} depending on the experiment enviroments.

In pair comparison, the probability of a subject preferring stimulus i over stimulus j can be defined as:

$$P(i;j) = \frac{q_i + d_i}{q_i + d_i + q_j + d_j} \tag{2.2}$$

Supposing there are n subjects, number of subjects k who selected i over j follows a binomial distribution with parameters n and $P(i;j)$. The Probability Mass Function (PMF) of the binomial distribution can be constructed as:

$$f(k;n,P(i;j)) = \binom{n}{k} \times P(i;j)^k \times (1 - P(i;j))^{n-k} \tag{2.3}$$

The optimal parameters q_i, d_{Home}, and d_{Lab} can be estimated by maximizing the

likelihood of the observed data. The likelihood function can be defined as:

$$L(P(i,j);k_{i,j},n_{i,j}) = \prod_{i<j} P(i;j)^{k_{i,j}}(1-P(i;j))^{n_{i,j}-k_{i,j}} \tag{2.4}$$

We can find the optimal parameters q_i, d_{Home}, and d_{Lab} by maximizing the likelihood function by using the matlab function proposed in [143].

The EBA model was initially designed for pairwise comparison subjective tests. However, our subjective test follows the DCR test, which does not involve pairwise comparisons. We construct the pairwise comparison matrix using the algorithm outlined in Algorithm 1.

The algorithm takes as input the DCR ratings provided by various observers under each condition. For example, if we are interested in comparing the impact of test conditions (AtHome *vs.* InLab), there will be only two conditions (C_1 and C_2). Nevertheless, this solution can be generalized to situations where more than two conditions are applied.

The algorithm first concatenates the DCR ratings for all conditions and then constructs the pairwise comparison matrix. The DCR rating for condition C_n is represented as a 2D list. Each row of C_n corresponds to different stimuli, while each column corresponds to different observers. The length of C_n is the number of stimuli tested in condition n. The number of observers who annotated stimulus s may vary across different conditions and stimuli.

The pairewise comparison matrix is a symmetric matrix, where the element PCM(i,j) is the number of subjects who prefer the stimulus i over the stimulus j. For a given observer o, if his/her DCR rating is higher for stimulus i than for stimulus j, we can infer that, in a pairwise comparison, he/she prefer stimulus i over stimulus j. However, in practice, asking the same observer to compare the same stimulus in different conditions is challenging, especially in crowdsourcing scenarios. Additionally, repeated voting for the same stimulus may introduce bias. Therefore, we address these challenges by randomly selecting the ratings from one observer for stimuli i and j and comparing them to construct the pairwise comparison matrix. The random shuffle in the algorithm is used to select random observers.

The algorithm is designed to handle the case where the randomly selected subjects' rating for stimulus i is equal to that of stimulus j. In this case, the algorithm will assign 0.5 to both PCM(i,j) and PCM(j,i).

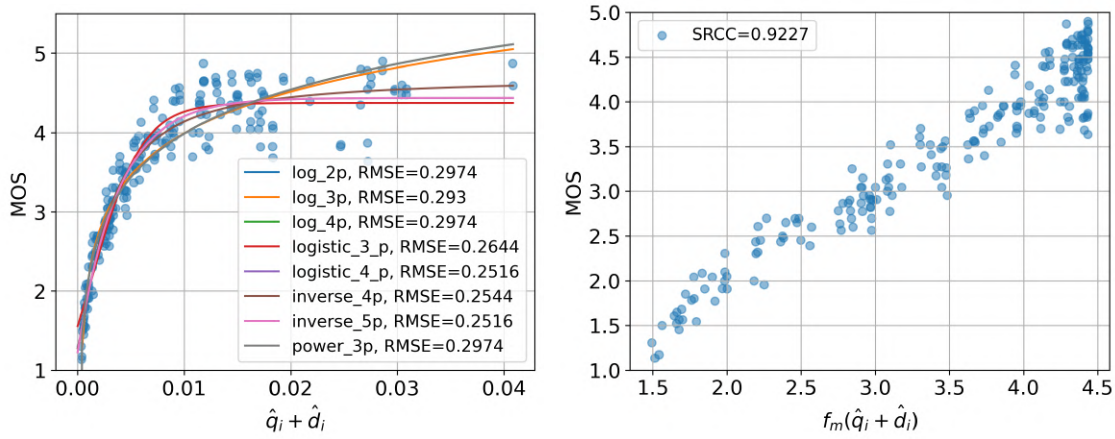
After maximizing the likelihood function (Eq.(2.4)), we can obtain the optimal param-

Algorithm 1 Pairwise Comparison Matrix Generation for EBA

Input: C_1, C_2, ...C_n - 2D list of DCR rating under every condition. Rows represent the stimuli and columns represent the observers.
Output: PCM - Pairwise Comparison Matrix.
function GeneratePCM(C_1, C_2, ...C_n)
    $C_{all} \leftarrow$ concatenate$(C_1, C_2, ...C_n)$
    $PCM \leftarrow$ zeros matrix of size $(len(C_{all}) * len(C_{all}))$
    for $i \leftarrow 0$ **to** $len(C_{all}) - 1$ **do**
        for $j \leftarrow 0$ **to** $len(C_{all}) - 1$ **do**
            if $i < j$ **then**
                $S_i \leftarrow C_{all}[i]$
                $S_j \leftarrow C_{all}[j]$
                $S_i \leftarrow$ random shuffle(S_i)
                $S_j \leftarrow$ random shuffle(S_j)
                $l_{min} \leftarrow \min(len(S_i), len(S_j))$
                for $obs \leftarrow 0$ **to** $l_{min} - 1$ **do**
                    if $S_i[obs] > S_j[obs]$ **then**
                        $PCM[i][j] \leftarrow PCM[i][j] + 1$
                    else if $S_i[obs] < S_j[obs]$ **then**
                        $PCM[j][i] \leftarrow PCM[j][i] + 1$
                    else
                        $PCM[i][j] \leftarrow PCM[i][j] + 0.5$
                        $PCM[j][i] \leftarrow PCM[j][i] + 0.5$
                    end if
                end for
            end if
        end for
    end for
    return PCM
end function

eters $\hat{q}_i$ for each stimulus and $\hat{d}_{\text{Home}}$, and $\hat{d}_{Lab}$ for each enviroment. We can then analyze the impact of the experiment environment on the QoE and quantify this impact.

Figure 2.8a shows the relationship between the sum of quality and experiment condition attributes and MOS. The MOS shows an increasing trend with the sum of the two attributes. We evaluated various mapping functions and found that the 4-parameter logistic function best represents the mapping function f_m between the sum of attributes and MOS. The Spearman Rank Order Correlation Coefficient (SROCC) between the mapped sum of attributes and MOS is 0.9227 (see Figure 2.8b). The SROCC value indicates a strong correlation between the mapped sum of attributes and MOS. However, it's worth



(a) Sum of quality and experiment condition attributes vs. MOS

(b) Logistic fitting of sum of quality and experiment condition attributes vs. MOS

Figure 2.8 – EBA analysis of the impact of experiment environment on the QoE

noting that the mapped sum of attributes tends to saturate for high-quality ranges.

EBA allows us to analyze the impact of different attributes separately. By eliminating the influence of the environment and retaining only the quality attribute of the stimuli, we can assess the impact of the experiment environment on MOS. As illustrated in Figure 2.9, for stimuli with the same quality attribute, the MOS of the AtHome experiment is comparable to that of the InLab experiment. This result further confirms that the AtHome and InLab environments do not have a significant impact on MOS, consistent with the conclusions drawn in Section 2.4.1 and Section 2.4.2.

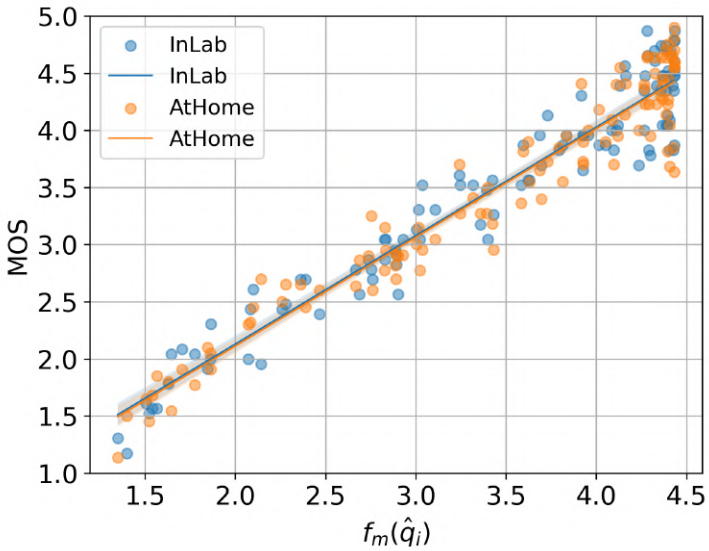


Figure 2.9 – Impact of experiment environment on the MOS

2.5 Display impact on subjective results

In the previous section, we demonstrated that the experiment environment has a limited impact on MOS (the observed quality for end users). In this section, we analyze the influence of the display on the MOS.

The HD DCR pre-qualification AtHome subjective test (see Section 2.3 for details) employed 10 TVs with 4 different TV models, as detailed in Table 2.1. Although the number of TVs varies for each TV model, *e.g.,* 4 SONY and 2 LG, and the number of observers is limited, *i.e.,* 2 observers per TV, we can still rely on the EBA model to gain a preliminary understanding of the impact of the display on the quality of experience for end users.

Similar to Section 2.4.3, we adapted the EBA model to measure the impact of the display on the MOS. The observed quality (MOS) is represented as $f_m(q_i + d_i)$ where q_i is the quality attribute and d_i is the display attribute. In our AtHome test, d_{SONY} represents the impact of the SONY displays; d_{LG} for the LG displays; d_{SG_TU} and d_{SG_Q74} for the 2 different SAMSUNG models, and the number of TVs for each TV models can be found in Table 2.1.

Negative impact of the display on the quality?

It is possible that some displays may have a negative impact on the quality. However, the EBA model assumes that the observed quality (MOS) is represented as $f_m(q_i + d_i)$. Does this mean that the EBA model assumes that the display can only have a positive impact on the quality? The answer is no. EBA assumes the values of attributes are **relative** and not in the MOS scale. For example, if the output of EBA is such that $d_{SONY} = 0.1$ and $d_{LG} = 0.05$, we cannot state that SONY improves the quality by 0.1 and LG improves the quality by 0.05. To determine whether the impact of the display is positive or negative, one should first choose one display as the reference and then compare the impact of the other displays to the reference. In the same example, if we choose SONY as the reference, then the impact of LG is -0.05, allowing us to conclude that LG has a negative impact on the quality compared to SONY.

As shown in Figure 2.10, we plot the impact of the display on the MOS for the 10 SRC of the AtHome pre-qualification test. The x-axis in each subfigure represents the normal-

ized mapped quality attribute. The same PVS has the same value of quality attribute. The y-axis represents the MOS. The color blue, orange, green and red correspond to SONY, LG, SG_TU and SG_Q74, respectively. It can be observed that for the same quality attribute, the MOS differs for different displays. This difference is due to the differences of the displays.

It can be observed that the impact of the display can vary between different SRC. For example, for SRC0, the impact of the display is relatively larger than for SRC3

From the fitted line with 1st order linear regression, we can observe that the LG display has the lowest MOS for the same quality attribute. Looking at the TV measuremets in Table 2.5, we can see that the peak luminance of LG is the highest among the 4 TV models. This result is consistent with the human contrast sensitive function(CSF) [27, 14, 94], which indicates that human visual sensitivity is higher for higher luminance, leading to a higher chance of detecting video distortions.

TV models	**SONY**	**LG**	**SG_TU**	**SG_Q74**
Peak luminance (cd/m²)	191.9960	249.9962	140.8884	232.6089

Table 2.5 – Peak luminance values for different TV models.

2.6 Summary

In this chapter, we introduced the "AtHome" subjective test system for conducting subjective tests "In-the-wild". The AtHome test system occupies a middle ground between the two widely used subjective test methods, InLab and Crowdsourcing. It effectively combines the controlled environment, such as display settings, of the InLab test with the diversity inherent in Crowdsourcing tests.

Furthermore, through an analysis of the experimental environment and display impact on subjective test results, we demonstrated the AtHome test system's ability to deliver results comparable to those obtained in the InLab environment. Additionally, the AtHome test system provided valuable insights into the influence of displays on subjective test outcomes. The variety and advanced functionalities of display ecosystems can significantly affect the Quality of Experience (QoE) of end-users.

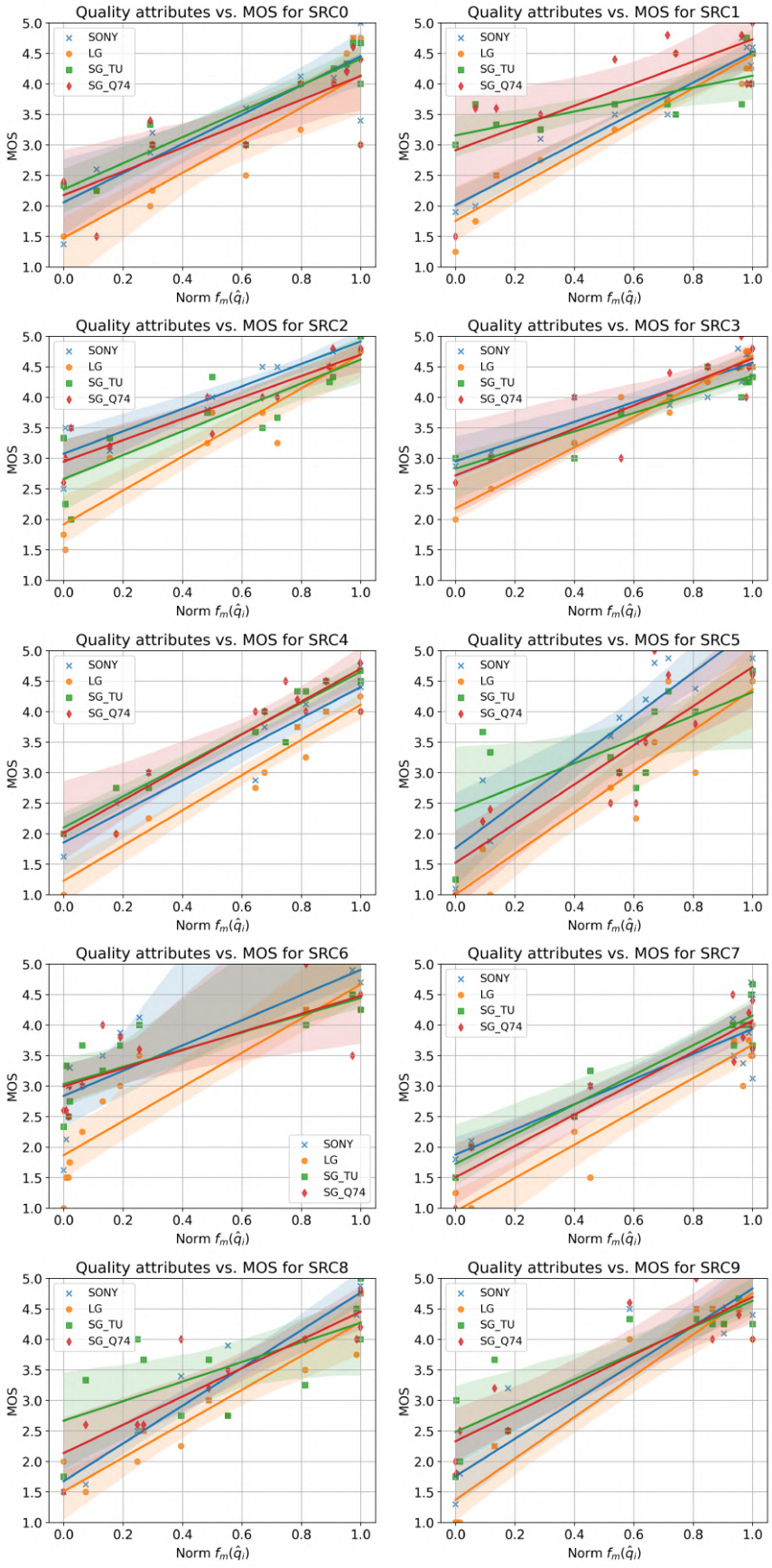


Figure 2.10 – Impact of the display on the MOS

Chapter Contributions

— Introduced a novel "AtHome" subjective test system to conduct subjective tests in a more ecological and accessible manner.

— Compared the AtHome test system with traditional in-lab subjective tests through pre-qualification tests, demonstrating its effectiveness and reliability.

— Analyzed the impact of different test environments and displays on the subjective test results.

Chapter 3

"In-the-Wild" Subjective Study of VW-JND

Overview

Contents

3.1 Introduction 31
3.2 Background and motivation 32
3.2.1 Subjective test for video quality assessment 33
3.2.2 Satisfied User Ratio (SUR) of JND 34
3.2.3 State of the art JND based datasets 37
3.2.4 Other relevant works 40
3.2.5 Motivation 40
3.3 JND search methodology 42
3.3.1 Related works 42
3.3.2 Simulation and comparison of JND search methods 44
3.3.3 Pre-processing of JCP 50
3.4 Content selection 52
3.5 Summary 54

Part of this chapter has been published in research papers [155, 154].

3.1 Introduction

Human eye cannot perceive small pixel changes in images or videos until a certain threshold of distortion. In the context of image/video compression, Just Noticeable Difference (JND) is the smallest distortion level from which the human eye just begins to perceive the difference between the anchor/reference stimuli and the distorted stimuli.

Picture-Wise Just Noticeable Difference (PW-JND) [90, 131, 126, 85], and Video-Wise Just Noticeable Difference (VW-JND) [137, 135, 151, 68] have been investigated from human perceptive aspects in order to provide **high quality** of experience for end-users with limited storage and internet bandwidth [153, text].

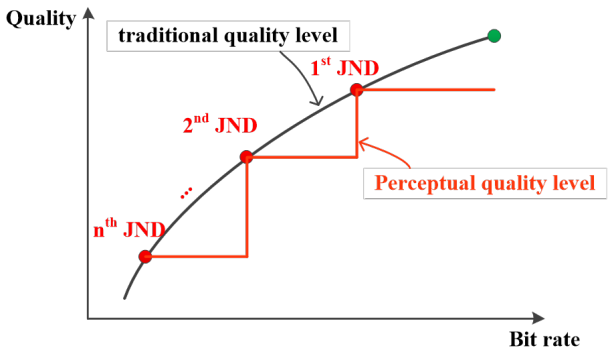


Figure 3.1 – Comparison of the traditional quality level and perceptual quality level on the Rate-Distortion curve. The traditional quality level is continuous, as widely used quality metrics such as PSNR and SSIM. In contrast, the perceptual quality level is discrete, based on thresholds beyond which the human visual system cannot distinguish differences.

Considering the limitation of publicly available VW-JND datasets, we have opted to gather new VW-JND datasets using the "In-the-wild" subjective test methodology outlined in the preceding chapter.

This chapter is organized as follows: In Section 3.2, we first introduce the background of the VW-JND study. We then formalize a more general SUR definition with different proxies in Section 3.2.2. Next, we compare the state-of-the-art JND datasets in Sections 3.2.3 and 3.2.4, and introduce the motivations behind collecting our new VW-JND datasets in Section 3.2.5.

In Section 3.3, we introduce different methods for JND search in the state-of-the-art and compare these methods using simulations. Moreover, we propose a pre-processing method to further optimize the JND search efficiency. In Section 3.4, we introduce the content selection process for our VW-JND datasets. Finally, Section 3.5 summarizes the chapter.

3.2 Background and motivation

In this section, we first introduce the background of the VW-JND study, including the subjective test for video quality assessment in Section 3.2.1. Secondly, we extend the

definition of the Satisfied User Ratio (SUR) of JND from encoding parameter as proxy to any distortion level as proxy in Section 3.2.2. Furthermore, we compare the state-of-the-art JND datasets in Section 3.2.3 and other related work in Section 3.2.4, and introduce the motivations behind collecting our new VW-JND datasets in Section 3.2.5.

3.2.1 Subjective test for video quality assessment

Subjective video quality tests are psychophysical experiments in which human subjects are asked to rate the quality of a video [58, 145, 23]. The goal of these tests is to measure the quality of experience (QoE) of the subjects, thereby aiding in the development of video processing algorithms such as video compression, enhancement, and restoration. Additionally, they serve as ground truth for the development of objective video quality assessment metrics.

Several subjective test methodologies for image and videos are standardized by the International Telecommunication Union (ITU) [56]. There are 4 main methods that are widely used today: Absolute Category Rating (ACR), Degradation Category Rating (DCR), Subjective Assessment Method for Video Quality (SAMVIQ), and Pair Comparison (PC) method.

ACR and DCR are both category rating methods since they use discrete scales, and they are dominant in video subjective quality tests [56, 127, 49]. The difference between ACR and DCR is that for ACR, observers are asked to rate the quality of a single stimulus, while for DCR, observers are asked to rate the quality degradation of a stimulus compared to a reference stimulus. DCR is also called Double Stimulus Impairment Scale (DSIS). The most widely used ACR and DCR rating scales are 5-point discrete scales, but there are also 3-point scales (usually used for Acceptance and Annoyance (AccAnn) [77]), 9-point, and 11-point discrete scales in various studies [54]. Moreover, some studies use 5-point/11-point continuous scales [53]

SAMVIQ [69] is more suitable for discriminating similar levels of quality. It is based on a random access process to play stimuli, allowing observers to start and stop the evaluation process, modify the quality score multiple times, and repeat the playback as they wish. It usually uses a 0-100 continuous scale. It is also called the multistimuli continuous quality scale in some studies [16, 112].

For the Pair Comparison (PC) method [58], test sequences are presented in pairs, usually side-by-side, and observers are forced to choose the one with the highest quality. PC usually combines all possible combinations, making it more time-consuming than

ACR and DCR. However, Mantiuk *et al.* [95] found that forced-choice PC method is more accurate than ACR and DCR.

The quality scores obtained from the aforementioned methods, such as Mean Opinion Score (MOS) and Differential Mean Opinion Score (DMOS) measure overall quality but do not provide information about JND. Similarly, widely used objective video quality metrics like VMAF [80], which are trained on these subjective datasets, reflect quality scores but are insufficient to determine if the HVS can perceive a difference between two videos. However, as shown in Figure 3.1, some small distortions may be imperceptible due to the psychological and physiological mechanisms of the Human Visual System (HVS) [32]. This presents an opportunity to optimize video transmission and storage without compromising perceptual quality for end-users.

For example, in HTTP Adaptive Streaming (HAS) [17, 8], video content is encoded at multiple bitrate-resolution pairs, known as **representations**, to construct the bitrate ladder. This allows for dynamic adjustment of video quality based on the viewer's available bandwidth and device type. By eliminating representations that are not perceptually different from the viewer's perspective, we can save the storage and bandwidth.

3.2.2 Satisfied User Ratio (SUR) of JND

For a given visual content, JND of different subjects will be different [87]. Wang *et al.*[136] proposed to conduct the subjective test of JND with respect to a viewer group other than the very few experts (golden eyes) for the worst-case analysis, because the group-based quality of experience (QoE) is closer to the realistic applications.

Satisfied User Ratio (SUR) curve can be derived from this group-based JND value. SUR curve is defined as the Q-function supposing that the group-based JND follows Gaussian distribution [137]. Intuitively, the value of SUR curve at a certain distortion level d, is the percentage of the group users who cannot perceive any difference between the reference stimuli and the distorted stimulus whose distortion level is smaller than d, *i.e.,* these users are satisfied.

At a given threshold p for SUR, the corresponding distortion level is defined as p%SUR instead of the misleading notation p%JND in previous works [137, 135, 151, 138].

The SUR value quantifies the portion of the population that cannot perceive distortion when a video is compared to a reference at a specific distortion level [138, 159]. This level is referred to as the **proxy** of the SUR curve. In literature, this proxy is usually encoding parameters such as QP (Quantization Parameter). Therefore, the SUR value decreases

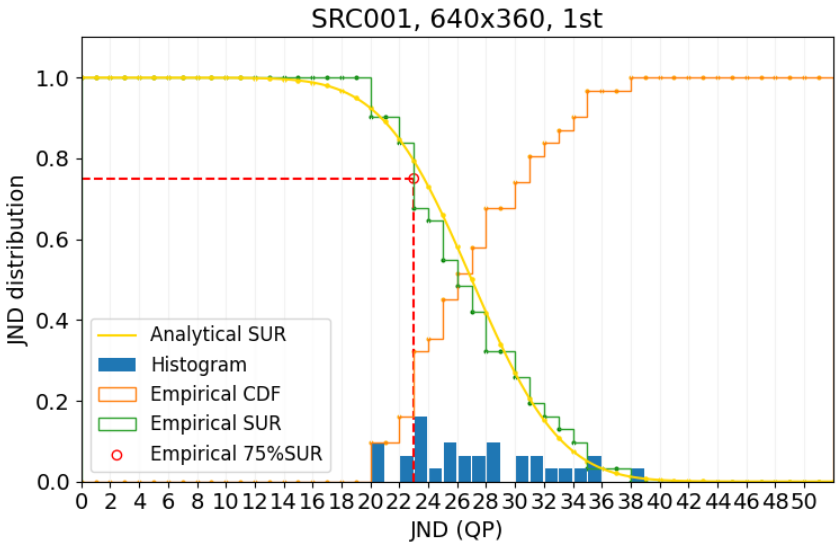


Figure 3.2 – Distribution of group-based VW-JND (blue bar); empirical and analytical SUR (in green and yellow respectively) and empirical 75%SUR (red circle) of SRC001 (360p) in VideoSet [138]

with the increase of the proxy value. In this section, **we extend the proxy of SUR to any metric that can reflect distortion level**, such as VMAF [80]. Unlike the QP proxy, for VMAF proxy, the SUR value will increase with the increase of the proxy value. We therefore define the SUR curve in two cases: where the quality decreases with an increase in the proxy (case 1) and where the quality increases with the proxy (case 2).

For a given video content clip m, let us assume there are VW-JND annotations from N reliable subjects. The VW-JND values from these N subjects can be represented as a vector $\mathbf{j^m}$, defined as:

$$\mathbf{j}^m = [j_1^m, j_2^m, ..., j_N^m] \tag{3.1}$$

Here, j_n^m represents the individual annotation of subject n, which can be QP or any other proxy capable of representing the distortion level, such as VMAF. Let J^m denote a discrete random variable representing the VW-JND for video m. $\mathbf{j}^m$ is a vector of random samples from J^m. The empirical Probability Mass Function (PMF) of J^m is given by:

$$p^m(x) = Pr(\text{JND} = x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(j_i^m = x\right), \tag{3.2}$$

where $\mathbf{1}(c)$ is an indicator function that equals to 1 if the specified binary clause c is true.

Thus, the empirical Cumulative Distribution Function (CDF) can be calculated from the PMF as follows:

$$\mathrm{CDF}^{\mathrm{m}}_{\mathrm{emp}}(x) = Pr(\text{VW-JND} \leq x) = \sum_{\omega<x} p^{m}(\omega). \tag{3.3}$$

In Figure 3.2, empirical CDF is represented in orange. Considering $\mathrm{SUR}_{\mathrm{emp}}$ to depend on the polarity of the chosen proxy, it is defined as follows:

$$\mathrm{SUR}_{\mathrm{emp}}(x) = \begin{cases} 1 - \mathrm{CDF}_{\mathrm{emp}}(x), \text{for case 1} \\ \mathrm{CDF}_{\mathrm{emp}}(x), \text{ for case 2} \end{cases} \tag{3.4}$$

In case 1, where quality decreases with an increase in the proxy (*e.g.,* using QP as the proxy as shown in Figure 3.2), the empirical SUR corresponds to the complementary empirical CDF. In contrast, in case 2, such as using VMAF as the proxy, where quality increases with the proxy increases, the empirical SUR is equals to the empirical CDF. Finally, $p\%\mathrm{SUR}_{\mathrm{emp}}$ is defined as:

$$p\%\mathrm{SUR}_{\mathrm{emp}} = \begin{cases} \min\left\{x \,|\mathrm{SUR}_{\mathrm{emp}}\left(x\right) \leqslant p\%\right\}, \text{for case 1,} \\ \max\left\{x \,|\mathrm{SUR}_{\mathrm{emp}}\left(x\right) \leqslant p\%\right\}, \text{for case 2.} \end{cases} \tag{3.5}$$

Figure 3.2 showcases the $75\%\mathrm{SUR}_{\mathrm{emp}}$ (represented by the red circle) for the QP proxy. We can determine $p\%\mathrm{SUR}_{\mathrm{emp}}$ for a specific video content using individual VW-JND annotations collected from a sampled population through subjective test.

The analytical SUR curve and p%SUR are calculated by Eq.(3.6) and (3.7), $f(x)$ is the Probability Density Function (PDF). Contrary to the empirical SUR, the analytical SUR is a continuous function, it can be obtained by fitting the empirical SUR curve with an assumption of the distribution of the individual VW-JND annotations (see Section 5.4.1 for more details). The analytical SUR curve is represented in yellow in Figure 3.2.

$$\mathrm{SUR}_{\mathrm{analy}}(x) = \begin{cases} 1 - \mathrm{CDF}_{\mathrm{analy}}(x), \text{for case 1} \\ \mathrm{CDF}_{\mathrm{analy}}(x), \text{ for case 2} \end{cases} \tag{3.6}$$

$$p\%\mathrm{SUR}_{\mathrm{analy}} = \arg\min_{x} |\mathrm{SUR}_{\mathrm{analy}}(x) - p\%| \tag{3.7}$$

The 75% SUR is the most widely used target prediction ground truth in previous works [137, 135, 152, 109, 108]. However, to our best knowledge, why the 75%SUR is selected as the target of prediction in previous works is not clear. In Chapter 6, we will investigate the

relationship between the SUR threshold and bitrate, thereby providing insight into how to decide the SUR threshold.

System bias for the SUR curve estimation?

In a recently published paper [63], the authors demonstrate that the SUR curve obtained by the previous definition introduces a system bias. An open discussion on this topic is available in Annex A. We approach this issue from a different perspective by introducing uncertainty estimation of the SUR curve in Section 4.3 of Chapter 4.

3.2.3 State of the art JND based datasets

There are several publicly available JND datasets for images and videos, as summarized in Table 3.1. For image JND datasets (PW-JND datasets):

— **MCL-JCI** [64] is a widely used image JND dataset. It includes 50 source images, each distorted using JPEG with Quality Factor (QF) values ranging from 1 to 100. Each source image is evaluated by 30 subjects, with over 150 participants in the subjective test. The reference image and its compressed version are displayed side by side on a 65-inch TV with a native resolution of 3840x2160. To efficiently update the comparison image, the binary search method is adopted [87]

— **JND-Pano** [92] is a panoramic image JND dataset consisting of 40 images with a resolution of 5000x2500. These images are compressed using JPEG with QF values ranging from 1 to 100. The subjective test is conducted in a lab environment using head-mounted displays (HMDs). Subjects have the flexibility to control the field of view (FoV) to explore the panoramic image. In contrast to MCL-JCI, the comparison images are displayed sequentially with the reference images.

— **SIAT-JSSI** [33] and **SIAT-JASI** [33] are stereo image JND datasets. The key distinction between the two datasets lies in their compression methods: SIAT-JSSI utilizes symmetric compression, while SIAT-JASI employs asymmetric compression. Both datasets incorporate two types of distortion: HEVC intra coding and JPEG2000. The subjective tests are conducted in a lab environment with a 3D display, where subjects wear polarized glasses to view the reference and distorted stereo images side by side. The relaxed binary search method [138] is utilized to update the comparison image.

Table 3.1 – Comparison of state-of-the-art JND-based datasets with our proposed datasets across various dimensions, including the year, content type, dataset size, number of subjects, resolution, distortion type, subjective test method, test environment, and JND search algorithm. The dataset size is detailed, indicating the number of pristine and distortion levels for each pristine. Similarly, the number of subjects includes both the count per content and the total number involved in the entire test.

Datasets name	Year	Content type	Size (Pristine/ distortion level)	Nb subjects (per content/ in total)	Resolution	Distortion type	Subjective test method	Test env	JND search algorithm
MCL-JCI [64]	2016	Image	50/ QF 1∼100	30/150	1920x1080	JPEG	PC side-by-side	Lab	Binary search
JND-Pano [92]	2018	Panoramic image	40/ QF 1∼100	25/42	5000x2500	JPEG	PC sequential	Lab	Binary search
SIAT-JSSI [33]	2019	Stereo image	12/ QP 1∼51	28/50	1920x1080	HEVC intra coding (symmetric)	PC side-by-side	Lab	Relaxed binary search
			12/ CR 1∼300	28/50	1920x1080	JPEG2000 (symmetric)			
SIAT-JASI [33]	2019	Stereo image	12/ QP 1∼51	28/50	1920x1080	HEVC intra coding (asymmetric)	PC side-by-side	Lab	Relaxed binary search
			12/ CR 1∼300	28/50	1920x1080	JPEG2000 (asymmetric)			
VVC [126]	2021	Image	202/ QP 13∼51	20/20	1920x1080	VVC	PC side-by-side	Lab	Binary search
KonJND-1k [84]	2022	Image	1008/ QF 1∼100	42/503	640x480	JPEG	Flicker (8Hz)	Crowd -sourcing	Slider
			1008/ QP1∼50	42/503	640x480	BPG			
MCL-JCV [136]	2016	Video	30/ QP1∼51	50/∼150	1920x1080	AVC	PC sequential	Lab	Binary search
Huang et al. [52]	2017	Video	40/ QP1∼51	30/30	1920x1080	HEVC	PC sequential	Lab	Binary search
VideoSet [138]	2017	Video	220/ QP 1∼51	30/800	1920x1080 1280x720 960x540 640x360	AVC	PC sequential	Lab	Relaxed binary search
FlickerVidSet [63]	2024	Video	45/ QP 0∼51	42∼48/51[1]	640x480	AVC	PC side-by-side +flicker	Crowd -sourcing	Quest+
			45/ QP 0∼63	41∼51/51[1]	640x480	VVC			
Our HD AMZ-HD-VJND	2022	Video	180/ dynamic CRF	20/20	1920x1080	HEVC	PC sequential	Home	Relaxed binary search
Our HDR AMZ-HDR-VJND	2023	Video	180/ dynamic CRF	20/20	3820x2160	HEVC	PC sequential	Home	Relaxed binary search

[1] Different from other methods, FlikerVidSet proposed a "Collective Observer" approach. For a given content, the common approach involves each observer conducting an entire JND search separately. However, for the "Collective Observer", each step of the JND search is conducted by a randomly selected observer.

— **VVC** [126] is similar to MCL-JCI, but it uses VVC compression instead. It offers a larger set of source images with fewer distortion levels. The subjective test is carried out in a controlled lab setting with a 55-inch smart TV.

— **KonJND-1k** [84] is a large-scale image JND dataset comprising 1008 source images compressed using JPEG and BPG with QF values ranging from 1 to 100 and QP ranging from 1 to 51. The subjective test is conducted in a crowd-sourcing environment. Unlike previous datasets, KonJND-1k employs a slider method instead of binary search to obtain the JND image. Additionally, instead of displaying the reference and distorted images side by side, they conduct a flicker test where the distorted image and the reference image alternate at a frequency of 8Hz.

For video JND datasets (VW-JND datasets):

— **MCL-JCV** [136] is a video JND dataset including 30 source videos compressed using AVC with QP values ranging from 1 to 51. Each source video is evaluated by 50 subjects, with a total of more than 150 participants in the subjective test. The reference video and its compressed version are displayed sequentially on a 65-inch TV with a native resolution of 3840x2160. The binary search method is used to update the comparison video.

— **Huang et al.** [52] collected a video JND dataset comprising 40 source videos. In contrast to MCL-JCV, the type of distortion used is HEVC instead of AVC.

— **VideoSet** [138] is a large-scale video JND dataset comprising 220 source videos compressed using AVC with QP values ranging from 1 to 51 for various resolutions (1080p, 720p, 540p, and 360p). Each source video is evaluated by more than 30 subjects, totaling over 800 participants. Relaxed binary search is used to track the JND threshold.

— **FlickerVidSet** [63] includes 45 selected and cropped sources from VideoSet. The subjective test is conducted in a crowd-sourcing environment. Instead of displaying the reference videos and the distorted versions sequentially, two side-by-side methodologies are used. In the first method, the reference video and the distorted video are played simultaneously side by side, referred to as the plain test. For the second method, the distorted video is replaced by a flicker video. This flicker video alternates frames of the reference video with frames of the distorted video at a temporal frequency of 8Hz. The Quest+ method is employed to track the JND threshold. They demonstrate that the flicker test is more sensitive than the plain test.

3.2.4 Other relevant works

Comprehensive JND Datasets in section 3.2.3 are all compression-oriented, meaning that the labels of JND points are based on levels of compression distortion. However, a generalized JND model should take different kinds of distortion into account. Therefore, Liu *et al.* [93] proposed a more comprehensive JND dataset for images. It contains 25 types of distortion, including Gaussian noise, motion blur, jitter, etc. Each distortion has 4 or 5 distortion levels. They used a coarse-to-fine strategy to select JND points, first collecting MOS (Mean Opinion Score) of distorted images as the preliminary selection criterion, and then using a flicker test to make fine JND selections. There are 106 source images and 1642 JND maps collected from 30 subjects using a crowdsourcing platform.

Zhang *et al.* [150] also collected a JND dataset to validate their Learned Perceptual Image Patch Similarity (LPIPS) metric. This dataset includes traditional distortions such as photometric distortion, compression distortion, and also CNN-based distortion. Each image patch among a total of 4.8k patches is distorted by one type of distortion randomly and evaluated by three subjects to determine if they can perceive the difference. It is worth noting that this dataset is not designed to track the threshold of JND distortion because only one type of distortion is applied to each image patch, and the distortion level is not varied.

2AFC JND Two-Alternative Forced Choice (2AFC) with scale reconstruction is also used to determine JND. Hoffman *et al.* [46] defines 1 JND as a difference in the image that, in a 2AFC test, observers can correctly identify the reference image between the reference and the distorted image with a probability of 75%. They collected a database of 18 images with more than 250k responses from observers using a flicker test. Pérez-Ortiz *et al.* [117] proposed a method to scale the 2AFC to a unified quality scale, such that a difference of 1 on the scale is equal to 1 JND.

3.2.5 Motivation

Despite the availability of existing datasets, several challenges remain to be addressed. Firstly, all current VW-JND datasets use Quantization Parameter (QP) as the distortion proxy. However, QP has limitations when applied to real-world industry settings, particularly in HTTP Adaptive Streaming (HAS).

This is because QP controls the level of compression but does not ensure a consistent bitrate, which is crucial for maintaining streaming quality in bandwidth-constrained en-

vironments. To tackle this issue, we collected datasets using Constant Rate Factor (CRF) as the distortion proxy, which is more practical for maintaining a relatively consistent bitrate compared to constant QP [124].

What is the link between p%SUR and 2AFC JND?

Giving one reference image/video and its distorted version, if the distance between these two is 1 JND according to a 2AFC test with scale reconstruction, then the SUR value between them is 50%. More generally, if the probability of correctly selecting the reference during a 2AFC test is c%, then it can be computed as:

$$c\% = \frac{1}{2}p\% + (1 - p\%) = 1 - \frac{1}{2}p\% \tag{3.8}$$

Where p%SUR means that $p\%$ of observers cannot perceive a difference between the reference and the distorted version; they will randomly vote during a 2AFC, while $1 - p\%$ of observers can perceive a difference, and they will vote for the reference during a 2AFC. Typically, for a 75%SUR, the corresponding c of 2AFC value is 62.5%, and for a 50%SUR, it's 75%.

Secondly, existing VW-JND datasets are primarily designed for Standard Dynamic Range (SDR) video, yet there is a growing demand for High Dynamic Range (HDR) videos due to the increasing availability of HDR contents [30, 110]. Therefore, we collected VW-JND datasets specifically for HDR content.

Thirdly, the resolution of existing datasets is typically limited to HD resolution. To address the need for high-quality content, we collected UHD resolution VW-JND datasets.

Moreover, most existing datasets are collected in lab environments, limiting their ecological validity. To overcome this limitation, we collected new VW-JND datasets in a home environment, as described in Chapter 2, to better reflect real-world viewing conditions.

Additionally, one of the major challenges in JND subjective testing is the time-consuming nature of the JND search process. For example, in the VideoSet dataset, each Source Video Sequence (SRC) is encoded into 51 Processed Video Sequences (PVSs) using H.264 with QP ranging from 1 to 51, leading to significant time costs. To mitigate this issue, we proposed preprocessing of JND candidate playlists (JCP) to incorporate a dynamic range of distortion levels instead of a fixed range (details are provided in Section 3.3.3). The comparison of our proposed JND datasets with existing datasets is summarized in

Table 3.1.

3.3 JND search methodology

In Section 3.2.3, we outlined the JND search algorithms used in current JND datasets. The goal of these algorithms is to determine the Just Noticeable Difference (JND) threshold by subjective tests. Typically, the JND search employs a psychophysical approach, where human observers assess and provide feedback on differences between the reference and distorted versions. This process demands considerable time and attention from participants. Hence, selecting an efficient JND search method is crucial for effective dataset collection.

In this section, we will introduce psychophysical methods applicable for JND searches (Section 3.3.1). Following this, we'll simulate and compare the accuracy and efficiency of various JND search methods (Section 3.3.2). Lastly, we'll introduce our proposed optimization of JND search methodologies through pre-processing of JCP in Section 3.3.3.

3.3.1 Related works

— **Method of Limits**, also known as the linear method, is a psychophysical method introduced by Fechner in 1860 [35]. It is a simple method that involves presenting stimuli in ascending or descending order of intensity until the observer can perceive the stimulus. The ascending and descending orders are repeated several times, and the results are averaged to obtain the threshold estimation.

— **Slider**: Lin *et al.* [84] proposed using a slider-based adjustment combined with the flicker test to collect a large-scale image JND dataset. Observers can freely drag the slider to adjust the distortion level of the distorted image until they can perceive the flicker effect. The authors claimed that the slider method can obtain comparable results with binary search while being more efficient. However, the slider method is more suitable for images than for videos.

— **Method of Constant Stimuli** presents stimuli in a quasi-random order that ensures each will occur equally often [31]. The observer is asked to respond whether they can detect the difference between the reference stimulus and the distorted stimulus. The threshold is estimated by fitting a psychometric function to the data. Each stimulus needs to be presented multiple times (usually not less than 20); therefore,

this method is rather time-consuming and requires patient, attentive observers.

— **Simple Staircase** is a modification of the Method of Limits. It is introduced by Békésy in 1960 [15]. The stimulus intensity is varied in a stepwise manner, and the observer is asked to respond whether they can detect the difference between the reference stimulus and the distorted stimulus. The step size is decreased after each reversal, and the threshold is estimated by averaging the reversal points. The problem with this Simple Staircase method is that the observer may easily become aware of the pattern of the stimulus intensity presentation, leading to biased threshold measurement. Therefore, there are many variations of the staircase method, such as the two interleaved staircase [26].

— **PEST**: Parameter Estimation by Sequential Testing (PEST) uses Maximum Likelihood Estimation (MLE) to select the most efficient stimulus intensity for a given trial [83]. Different with the staircase method, the PEST method can adaptively adjust the step size according to the observer's response. It usually assumes the psychophysical function is a sigmoid function.

— **Quest+** is a Bayesian adaptive psychometric method that estimates the threshold by fitting a psychometric function to the data [142]. It is a generalization of the original QUEST method [141]. It is an adaptive variant of the Method of Constant Stimuli, and it can adjust the stimulus intensity adaptively according to the observer's response. Mohsen *et al.* [63] adopted Quest+ to collect VW-JND datasets through crowdsourcing.

— **Binary Search** is a widely used algorithm in computer science. Taking the JND Candidate Playlist (JCP) with QP from 0 to 51 with step size 1 as an example (see Figure 3.3 left), the observer will be asked initially if they can perceive the difference between the video with QP = 0 (Reference) and QP = 25 (middle of the original interval of JCP). If "YES", the interval QP = [26, 51] will be excluded in the next comparison; if "NO", the interval QP = [0, 24] will be excluded.

— **Relaxed Binary Search**: Binary Search may encounter issues when the observer makes an unconfident decision in the previous comparison. The Relaxed Binary Search, proposed by Wang *et al.* [138], is a modified version of Binary Search. It only eliminates one quarter of the original interval instead of half, for example, interval QP = [39, 51] instead of QP = [26, 51] in the previous example (see Figure 3.3 right).

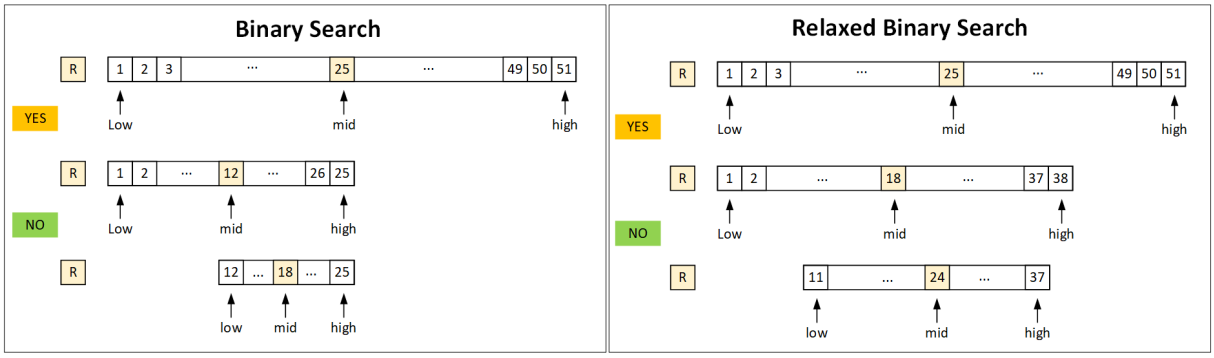


Figure 3.3 – Binary Search vs. Relaxed Binary Search: R stands for reference, the list on the right is the JCP with QP from 0 to 51 with step size 1. yes and no indicate the observer's response.

3.3.2 Simulation and comparison of JND search methods

In this section, we aim to compare the accuracy and efficiency of various JND search methods to understand the pros and cons of these methods. We will be examining three widely used JND search methods: Relaxed Binary Search, Simple Staircase, and Quest+. The comparison will be based on both experiment time efficiency and JND measurement accuracy.

Experiment time efficiency will be quantified by the number of trials or comparisons required, while JND measurement accuracy will be evaluated using the Mean Absolute Error (MAE) between the measured JND and the ground truth JND of the observer simulation model.

Observer simulation model

We employed the same assumption as [63, 138] that for a given content, observers' JND thresholds follow a Gaussian distribution with mean μ and standard deviation σ. Here, μ represents the sensitivity to distortion of each observer, while σ indicates the consistency of the observers' responses. In Figure 3.4, we showcase three different observers. Observer 1 is the most sensitive to distortion, while observer 2 is the least sensitive to distortion but the most consistent in response. To create a more realistic observer model, we referred to the publicly available VW-JND dataset VideoSet [138]. Among the 4 resolutions available in VideoSet, we only use the 1080p resolution as reference. It was observed that for the 220 Source Content (SRC) videos in VideoSet 1080p, the range of μ and σ are [15, 35] and [2.36, 9.38] respectively. These values were used to conduct the following simulations.

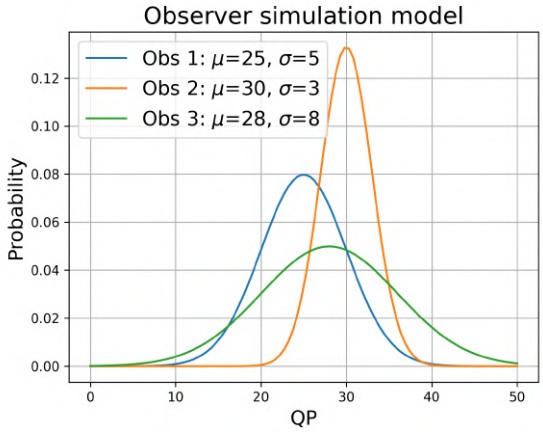


Figure 3.4 – Observer simulation model

Experiment time

Subjective tests to determine the JND threshold are considerably longer than other classic subjective tests for image and video quality, such as ACR and DCR tests. Given budget constraints and observer fatigue concerns, it is crucial to measure the experiment time of each JND search method. The subjective test time is determined by the number of trials or comparisons required.

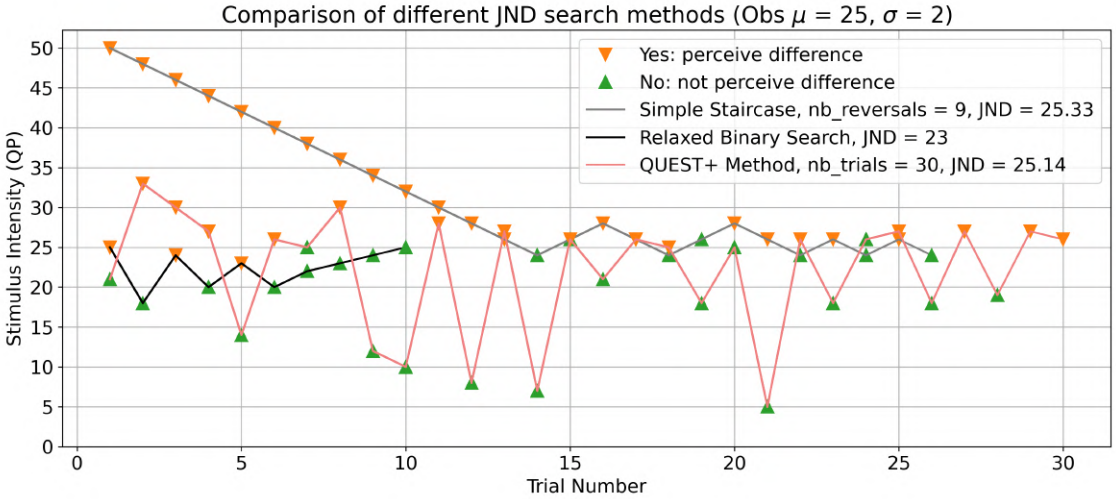


Figure 3.5 – Comparison of Relaxed Binary Search, Simple Staircase, and Quest+ in terms of the number of trails/comparison for the observer model follows $\mathcal{N}(25, 2^2)$

Simple Staircase stops the search when the number of reversals reaches a predefined

number, typically set to 6 to 9 [31]. The number of trials is determined by the number of reversals. In Figure 3.5, the gray line represents the simulation process of the simple staircase method with descending order initialization and a step size of 2. In this example, the number of reversals is 9, resulting in 26 trials. The final JND is 25.33, with an MAE of 0.33 compared to the ground truth JND of the observer model.

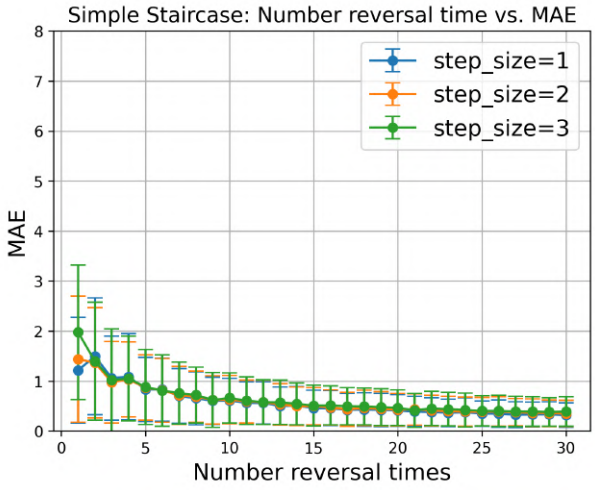


Figure 3.6 – Number of reversal vs. MAE for Simple Staircase

We furthermore conducted simulations to understand the relationship between the number of reversals for stopping the simple staircase search and the MAE of the obtained JND. As shown in Figure 3.6, the MAE decreases as the number of reversals increases for different step sizes. The simulations were conducted for observers following a $\mathcal{N}(25, 2^2)$ distribution, repeating 1000 times. Additionally, we conducted simulations for different observer models, and the results can be found in Annex B. The MAE increases with the increase of the standard deviation of the observer model, while the mean of the observer model has almost no impact on the MAE.

Interestingly, the step size impacts the MAE differently for different standard deviations of observer models. For higher observer consistency (*i.e.*, smaller σ), smaller step sizes perform slightly better than larger step sizes. However, for lower observer consistency (*i.e.*, larger σ), larger step sizes perform significantly better than smaller step sizes. This indicates that a larger step size is more robust for inconsistent observers (details can be found in Figure B.3 in Annex B).

Relaxed Binary Search will stop the search when the search interval is less than 2 [138]. For classic binary search, it is well known that the maximum number of comparison

is

$$\log_2 \left(len(JCP) \right) = \log_{(1/2)} \left(1/len(JCP) \right), \tag{3.9}$$

However, the interval to keep for relaxed binary search is 3/4 instead of 1/2 in each iteration, thus the maximum number of comparison is calculated by $\log_{(3/4)} \left(1/len(JCP) \right)$. Figure 3.7 shows the number of trials required for the relaxed binary search method with different JCP lengths. The simulation was conducted for various observer models. The number of trials increases as the length of the JCP increases, following a logarithmic relationship. For VideoSet [138], where the length of the JCP is 51, the number of trials ranges between 10 and 11 for different observer models. These simulation results are consistent with those reported in [63].

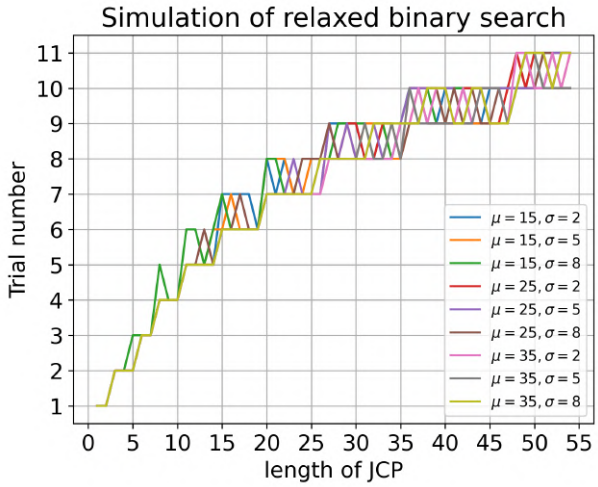


Figure 3.7 – Number of trials vs. length of JCP for Relaxed Binary Search

For **Quest+**, Jenadeleh *et al.* [63] set the trial number to 30, while Parie *et al.* [116] used 64-trials Quset+ procedure to determine JND. We conducted simulations for Quest+ [1] to explore the relationship between the number of trials and the MAE of JND measurement.

As shown in Figure 3.8, the MAE decreases as the number of trials increases for observer model following a $\mathcal{N}(25, 2^2)$ distribution repeated 1000 times. We also conducted simulation for different observer models and the results can be found in Annex C.

Similar to the Simple Staircase method, the MAE increases with the standard deviation of the observer model. Across different observer models, the MAE decreases with

1. https://github.com/hoechenberger/questplus.

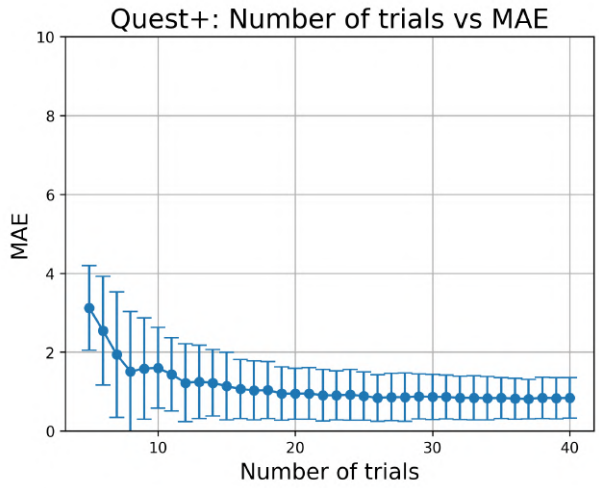


Figure 3.8 – Number of trials vs. MAE for Quest+

an increase in the number of trials. The comparison of the three JND search methods in terms of the number of trials and MAE will be further discussed in Section 3.3.2.

JND measurement accuracy

We simulate the JND search process for different observer models and compare the accuracy of the JND measurement for the three JND search methods: Simple Staircase, Relaxed Binary Search, and Quest+. Detailed results can be found in Annex D. We used the same JCP as VideoSet (QP 1 51) for simulation, and μ ranges from 15 to 35 and σ ranges from 2 to 8 for the observer models. The MAE is averaged over 1000 simulations for each observer model.

For **Simple Staircase**, we varied the reversal times (from 6 to 8) as the stop condition, each with step sizes ranging from 1 to 3. Observations from Table D.1 in Annex D are as follows:

— Increasing the step size leads to lower trial numbers. However, the impact of step size on the Mean Absolute Error (MAE) varies across different observer models. For observers with higher consistency, smaller step sizes perform slightly better, whereas for less consistent observers, larger step sizes are significantly better, as also shown in Figure B.3 of Annex B.

— Changes in the μ value do not significantly affect the MAE, while changes in the σ value have a notable impact. MAE increases with higher standard deviations of the observer model for different Simple Staircase settings.

— The μ value significantly impacts the trial numbers. Higher μ values lead to lower trial numbers, as the staircase begins from the highest distortion level in descending order. Conversely, if we set the staircase to begin from the lowest distortion level in ascending order, the trial numbers will increase with higher μ values.

— For the same step size, increasing the reversal time for the stop condition will result in higher trial numbers but also improve the MAE. The MAE decreases with the increase of the reversal time for different observer models.

For **Relaxed Binary Search**, observations from Table D.1 in Annex D are as follows:

— Trial numbers always range between 10 to 11 for different observer models with a fixed JCP length (consistent with Figure 3.7).

— Similar to Simple Staircase, the accuracy of JND measurement decreases with higher standard deviations of the observer model, while the impact of the mean is not significant.

— To achieve a similar MAE as the Simple Staircase method, Relaxed Binary Search requires fewer trials, making it more efficient in terms of experiment time.

For **Quest+**, observations from Table D.1 are as follows:

— MAE decreases with an increase in the number of trials for different observer models.

— Similar with Simple staircase and Relaxed Binary Search, the JND measurement accuracy reduces with the increase of the standard deviation of the observer model.

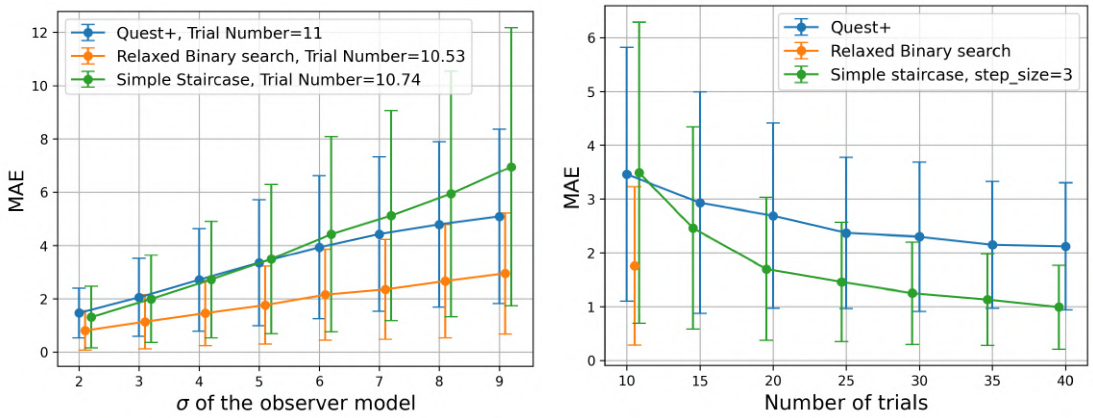
— Similar to Simple Staircase and Relaxed Binary Search, JND measurement accuracy decreases with higher standard deviations of the observer model.

— With the same trial numbers as Relaxed Binary Search, the accuracy of Quest+ is significantly lower compared to Relaxed Binary Search.

If we maintain consistent trial numbers for the three JND search methods mentioned above, we can plot the MAE of each method with different observer models. As depicted in Figure 3.9a, the higher the inconsistency of the observer model, the greater the MAE for all three methods. Notably, the Relaxed Binary Search method outperforms the other two methods in terms of MAE across various observer models.

The trial number for Relaxed Binary Search is only determined by the length of the JCP, whereas the trial numbers for Simple Staircase and Quest+ are predefined. As illustrated in Figure 3.9b, with a fixed observer model, the MAE decreases as the trial numbers increase for both Simple Staircase and Quest+. While the Simple Staircase and

Quest+ methods can achieve higher accuracy than the Relaxed Binary Search method, they require significantly more trials. For instance, Simple Staircase needs twice as many trials as Relaxed Binary Search to attain the same MAE.



(a) MAE vs. observer σ (b) MAE vs. trial numbers

Figure 3.9 – Comparison of different JND search methods: (a) Maintaining nearly equal trial numbers and observer $\mu = 25$, MAE versus observer σ for different JND search methods. (b) Keeping the same observer model (μ=25, σ=5), MAE versus trial numbers for different methods.

In conclusion, the Relaxed Binary Search method proves to be more efficient in terms of experiment time compared to Simple Staircase and Quest+. Although Simple Staircase and Quest+ can achieve higher accuracy than Relaxed Binary Search, they necessitate significantly more trials. Considering the need for collecting large-scale VW-JND datasets and budget limitations, the Relaxed Binary Search is a more practical choice for collecting our HD and HDR VW-JND datasets.

3.3.3 Pre-processing of JCP

The JND search procedure remains time consuming, even with the fastest search method: relaxed binary search. For example, for a JCP lenght of 51 in VideoSet (QP 1 51), it requires 10 to 11 trials to find the JND (details are provided in Section 3.3.2). The total time for one observer to find the JND for one content T_{jnd} can be computed as:

$$T_{jnd} = (t_{video} \times 2 + t_{rating}) \times N_{tiral} \tag{3.10}$$

where t_{video} is the duration of the video sequence, multiplied by 2 because the videos are played sequentially. t_{rating} is the duration of the rating, and N_{trial} is the number of trials. For a 10-second video and a 5-second rating, the total time for one observer to find the JND for one content is more than 4 minutes, which is still challenging for collecting a large-scale VW-JND dataset.

The JND search time depends on the length of JCP. The longer the JCP is, the longer the search time will be for different JND search methods. Meanwhile, it is well known that from a certain level of compression (e.g., QP = 40), it is almost certain that anyone with correct visual acuity can perceive the difference between the reference and the PVS.

Therefore, we proposed a method to optimize the JND subjective test time by reducing the length of JCP with the help of a pre-processing using the mapping function from VMAF to JND proposed in [155]. The mapping function for HD videos is shown in Figure 3.10. It can be observed that the higher the VMAF difference (ΔVMAF) between two videos (same content, different encoding recipes), the more likely it is for humans to perceive differences between them in terms of quality.

The idea is to remove the low quality PVSs that human eyes can perceive "for sure" differences to reduce the numbers of comparison before finding the JND. For a given threshold $thr\%$, the corresponding value of ΔVMAF in the mapping function is denoted as $V_{thr\%}$. The reference for the *1st* JND is SRC, therefore $\Delta\text{VMAF} = \text{VMAF}(SRC) - \text{VMAF}(PVS)$ and $\text{VMAF}(SRC) = 100$. The PVS whose ΔVMAF is larger than $V_{thr\%}$ will be removed from the JCP. Eq.(3.11) stipulates the condition to eliminate the PVS to save subjective test time.

$$\text{VMAF}(PVS) < 100 - V_{thr\%} \tag{3.11}$$

As shown in Table 3.2, we compared the trial numbers and experiment times. The "Mean of $len(JCP)$" represents the average length of the JCP across the entire datasets. As the threshold $thr\%$ in the mapping function decreases, $V_{thr\%}$ decreases and the number of PVSs eliminated increases according to Eq.(3.11), thus decreasing the average length of JCP. However, there's a possibility that the JND video may also be excluded during this procedure. The last columns in Table 3.2 indicate the number of videos whose JNDs are excluded during the pre-processing. It can be observed that only when the threshold is close to 1, we can ensure not to remove any JND in VideoSet.

The maximum number of comparisons/trials is calculated as in Section 3.3.2, and the duration is estimated by Eq.(3.10). It can be concluded that our proposed method can

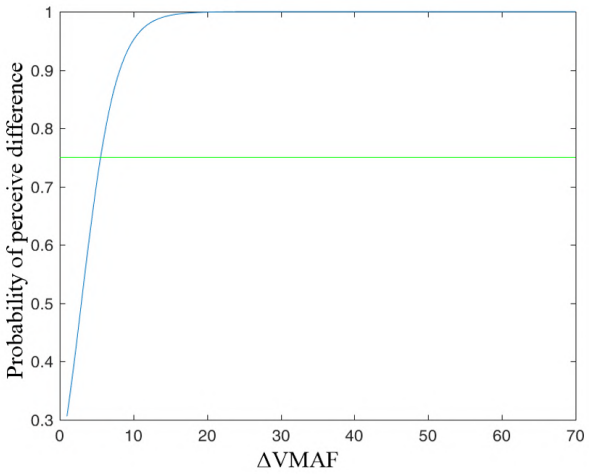


Figure 3.10 – Mapping function between ΔVMAF and probability of perceive difference between 2 videos with different encoding recipes for HD videos

reduce the subjective test duration by 9.09% without removing any mandatory information.

Table 3.2 – Benchmark between our solutions and original relaxed binary search in terms of testing time in VideoSet (1080p)

JND search method		Mean of *len*(JCP)	Max trial number	Duration (s)	JND excluded
baseline[138]		52	11	275	0
thr. =	99%	36.99	10	250	0
	95%	29.54	9	225	45
	85%	27.21	9	225	113
	75%	24.85	8	200	159

3.4 Content selection

The videos for our JND datasets were provided by the Prime Video team. These source videos cover a wide variety of genres of Video on Demand (VOD), including action, drama, comedy, TV shows, documentaries, animations, etc. All the videos are cut into 10-second-long sequences with a constraint on the number of scene cuts. Due to budget limitations, we have decided to collect one HD SDR and one UHD HDR VW-JND dataset, each containing 180 video contents. To ensure a comprehensive coverage of content types, we need to select representative contents.

The considered features to characterize the contents are:

— **Spatial Information (SI)** [55] reflects the spatial complexity of videos. Higher SI values indicate more details, contrast, and edges in the videos.

— **Temporal Information (TI)** [55] reflects the temporal complexity of videos. Higher TI values indicate more motion and dynamics in the videos.

— **Colorfulness** is an important visual feature [4]. We used the metric proposed by Hasler *et al.*[44] to compute the colorfulness of the videos.

— **Texture features** can be extracted by computing the Gray Level Co-occurrence Matrix (GLCM) [42]. GLCM is, in fact, a 2-D histogram given distance and angle.

— **Ambiguity** is a measure of the difficulty observers face in judging content quality. We used the content ambiguity features proposed by Ling *et al.* [88], which are based on content ambiguity derived from subjective test results proposed by Li *et al.* [79] as training labels.

— **Bitrate-Distortion cluster** is a feature to characterize the behavior of video content towards compression. Ling *et al.* [89] proposed a Bitrate-Distortion rate clustering method to classify content into different clusters based on the slope of the BD-rate.

For HDR videos, besides the above features, we also considered the following features:

— **MaxCLL** (Maximum Content Light Level): This represents the maximum brightness level in the video.

— **MaxFALL** (Maximum Frame Average Light Level): This indicates the maximum average brightness level in the video.

— **WCG** (Wide Color Gamut): These features describe the color volume of the video. We used the metric proposed by Lee *et al.* [73, 74].

— **HDR contrast**: We used the features proposed by Narwaria *et al.* [111] for HDR image content selection. These features reflect the quality difference between the original HDR content and contrast-reduced content.

We compute these features for all the 10-second video clips. For WCG and HDR contrast, we only compute the features for the first frame of the video clips. These features are then input into a K-means clustering algorithm to group the videos into different clusters. We select videos from each cluster to ensure a broad coverage of content types.

3.5 Summary

In this chapter, we first introduce the concept of JND and its significance in video quality assessment. Subsequently, we generalize the definition of SUR with a proxy applicable to any distortion level. SUR stands as one of the most crucial concepts in this thesis. Following that, we discuss the motivation behind collecting new VW-JND datasets after comparing the currently available JND datasets across different aspects.

Next, we delve into the JND search methods, which constitutes the most time-consuming part of the JND subjective test. We compare the accuracy and efficiency of three widely used JND search methods: Relaxed Binary Search, Simple Staircase, and Quest+. Our findings reveal that the Relaxed Binary Search method is more efficient in terms of experiment time compared to Simple Staircase and Quest+. Based on the simulation results, we propose a pre-processing method to optimize the time efficiency of the JND subjective test.

Finally, we provide a brief overview of the content selection process for our VW-JND datasets. The new HD and HDR datasets, named AMZ-HD-VJND and AMZ-HDR-VJND respectively, were collected following the subjective test methodology described in this chapter and the "In-the-Wild" subjective test environment set-up in Chapter 2. Further analysis of the collected dataset will be presented in the next chapter.

Chapter Contributions

- Extended the definition of SUR to include proxies beyond encoding parameters.
- Benchmarked various JND search methods, evaluating both their accuracy and efficiency.
- Proposed a pre-processing method to significantly enhance the time efficiency of JND subjective tests.

Chapter 4

Subjective Data Analysis

Overview

Contents

4.1 Introduction 55

4.2 ZREC: robust recovery of mean and percentile opinion scores 56

4.2.1 Background and motivation 56

4.2.2 QoE Datasets 59

4.2.3 Proposed Model 59

4.2.4 Experiment Results 63

4.2.5 Conclusion 65

4.3 Uncertainty analyses of SUR 66

4.3.1 Motivation 66

4.3.2 Uncertainty estimation of p%SUR$_{\text{emp}}$ 67

4.3.3 Uncertainty estimation of SUR curve 71

4.4 Longitudinal study of Subjective Data 75

4.4.1 Test campaign management 75

4.4.2 Observer behavior analysis 77

4.5 Summary 81

Part of this chapter has been published in research papers [161, 156].

4.1 Introduction

After collecting subjective datasets, it's crucial to analyze the data before developing any objective prediction models or metrics. This analysis helps determine the reliability of the collected data and identify any outliers that may need to be addressed. To address

these concerns, we propose a new data screening method called ZREC in Section 4.2.

Furthermore, while uncertainty has been widely studied for traditional rating experiments, there has been limited analysis of uncertainty for Satisfied User Ratio (SUR) of JND. In Section 4.3, we present an uncertainty analysis for the SUR curve. This analysis is crucial when using the SUR for further analysis or model training.

Additionally, the subjective tests of the VW-JND pipeline introduced in previous chapters (2, 3) present an opportunity to conduct a longitudinal study of the subjective data. Unlike classic In-Lab subjective tests, our At-Home subjective tests were conducted over a longer period. In Section 4.4, we present the longitudinal analysis of the subjective data.

4.2 ZREC: robust recovery of mean and percentile opinion scores

Observer screening and subject opinion score recovery is essential for collecting a reliable QoE database. In this section, we propose a new method, ZREC [1], which uses Z-scores to estimate subject bias, inconsistency, and content ambiguity. Additionally, we propose Mean Opinion Score (MOS) recovery and Percentile Opinion Score (POS) recovery scheme based on the three estimated parameters. ZREC does not fully reject subjects, rather adjust their coefficients in the MOS/POS recovery, allowing for more efficient use of data collection. The estimated parameters of ZREC are highly correlated with more complex solver-based methods and standards. In addition, ZREC recovers MOS with smaller confidence intervals than the state of the art. Experimental results also demonstrate that using recovered p_{th} POS as ground truth during training improves the performance of SUR prediction.

4.2.1 Background and motivation

Observer screening is an essential steps of collecting a reliable Quality of Experience (QoE) database. A range of methods [79, 75, 76, 2, 1] with varying complexities have been proposed, and various standards [56, 58, 57] include recommendations for this purpose. Observer screening can reduce personal bias, eliminate outliers, and provide higher

1. available at: https://github.com/kyillene/ZREC

confidence data. Additionally, it has also been demonstrated that training learning-based metrics on recovered MOS can enhance their performance to a certain extent [118].

The collected QoE measurements in subjective studies are often characterized as a combination of subject bias, inconsistency and the underlying quality of the stimuli [79]. As shown in Eq.(4.1), we model the raw opinion scores as a random variable $o_{i,j}$ for subjects i and stimulus j, it can be represented as the sum of the true opinion score O_j, the subject bias B_i, and the content ambiguity $A_j\varepsilon_j$, where ε_j is standard normal variable.

$$o_{i,j} = O_j + B_i + A_j\varepsilon_j \tag{4.1}$$

Subject bias refers to the systematic error of a subject towards a certain direction, *e.g.* a positive bias indicates the subjects overall tendency to perceive a higher quality. **Subject inconsistency** is associated with the random unexplained error included in the observations, such as lack of attention, malicious intentions etc. On another front, the **Content ambiguity** defines the level of difficulty in evaluating a stimulus due to its inherent ambiguity.

Below, we provide a summary of commonly used MOS recovery methods from the literature and briefly discuss their advantages and disadvantages. In addition, Table 4.1 provides a quick overview of the parameters estimated by each model.

Table 4.1 – Summary of the estimated parameters by each mos recovery model.

	BT500	P913	P910	MLE	ZREC
Subject Inconsistency	✗	✗	✓	✓	✓
Subject Bias	✗	✓	✓	✓	✓
Content Ambiguity	✗	✗	✗	✓	✓

— **BT500:** ITU-R BT.500 Recommendation [56] defines **outlier rejection** procedures. We use the widely adopted kurtosis-based outlier rejection procedure in BT.500 [56], where subjects are rejected based on the number of opinion scores outside of the predefined amount of standard deviation range of the population. If a subject found to be an outlier, all of his/her opinions are removed from the dataset. MOS is calculated as the mean of remaining subjects. Due to hard-coded thresholds and removing all votes of a detected outlier, the MOS recovery may result in even larger confidence intervals

— **P913:** also called **P913-12.4**. ITU-R P.913 Recommendation clause 12.4 [57] pro-

poses a procedure based on both **bias removal** and **outlier rejection**. For each subject, the bias is calculated as the average difference between the MOS and subjects' opinion score of each stimulus. Estimated biases are removed from subject opinion scores and then MOS can be calculated as the average of bias-removed opinion scores, optionally after rejecting outliers. It can be seen that P913 is a slight improvement over BT500. However after removing the subject bias, the observer are still rejected with hard-coded parameters or treated equally by ignoring the subject inconsistency.

— **P910:** also called **P913-12.6**. ITU-R P.913 Recommendation clause 12.6 [57] defines a procedure where MOS is recovered by **bias removal** and **subject inconsistency** weighting. Same procedure is also included in ITU-R P.910 Recommendation Annex-E [58]. The procedure defines the individual opinion scores of a subject as the combination of subject bias, inconsistency and the true quality of the stimuli and jointly solves these three parameters. Two solvers are proposed for the approach in [79]. Due to minimal differences between the solvers, we only consider the Alternating Projection (AP) solver in this work. P910 can be seen as the next step of the P913 by additionally considering subject inconsistency during MOS recovery. Note that the model does not provide any estimate for content ambiguity.

— **MLE:** Li *et al.* [78] proposed a a MOS recovery approach by jointly estimating bias and inconsistency of subject and content ambiguity with Maximum Likelihood Estimation (MLE) and belief propagation. In addition to bias and inconsistency of subject as P910, MLE also provide content ambiguity. However, it is acknowledged by the authors that the MLE solver has the issue of lacking of uniqueness in its solution in certain cases, *e.g.,* it cannot find solutions for our collected AMZ-HD-VJND dataset (see Section 4.2.2).

To address the limitations of previous work, we propose an alternative method that relies on Z-score to estimate subject bias, inconsistency, and content ambiguity. We also present a simple yet efficient MOS and POS recovery scheme. Our proposed model is more robust to different use-cases and datasets as it does not require a solver, which can sometimes result in convergence issues. The contributions of Section 4.2 are:

— A simple yet robust statistical model for estimating subject bias, inconsistency, and content ambiguity from subjective opinion scores.

— A MOS and POS recovery method based on the estimated subject bias and incon-

sistency.

— Performance comparison between the proposed model and the state-of-the-art, validated by estimating CIs and quantifying the impact of p_{th} POS recovery on the accuracy of SUR prediction models.

4.2.2 QoE Datasets

The performance of ZREC and other existing models from the literature in terms of MOS and POS recovery, as well as estimation of subject bias, inconsistency and content ambiguity, have been assessed on two datasets with distinct characteristics.

AMZ-HD-VJND: is our collected dataset for HD videos. There are 180 source content (SRC) evaluated by 20 naive subjects with correct visual acuity. Each SRC has been compressed with HEVC with different Constant Rate Factor (CRF) and presented to each subject via Relaxed Binary Search [154] to find the JND of each subject. Therefore the proxy of JND is represented by CRF value. SUR curve is the complementary cumulative distribution function of the individual JNDs of a viewer group [159]. q%SUR is the CRF value that corresponds to a SUR value on the SUR curve equals to threshold q%. 75% is the most commonly used threshold [151, 137, 152, 86, 34]. More detials about this datasets please refer to Chapter 2 and 3.

In this work, the individual JND annotations for each subject are considered as the opinion scores. Additionally, the q%SUR is equivalent to the p_{th} percentile ($p = 1 - q$, see Eq.(4.12) in Section 4.2.3) of opinion score. The opinion scores were used for MOS/CI validation and parameter estimation experiments as well as to measure the impact of POS recovery on the accuracy of SUR prediction models.

Netflix Public: Netflix Public Dataset [113] is a publicly available video quality dataset with 79 Processed Video Sequences (PVS) where each evaluated by 26 subjects. We used the opinion scores for MOS/CI validation and parameter estimation experiments.

4.2.3 Proposed Model

Let $o_{i,j}$ be the opinion score annotated by subject i for stimulus j. For a subjective dataset that consist of m stimulus and have been evaluated by n subjects, the original annotation can be represented by a matrix $\mathbf{O} \in \mathbb{R}^{n \times m}$. For every stimulus, we first compute the mean and standard deviation of the opinion score annotated by each subject:

$$\mathbf{m}(j) = \left(\frac{1}{n} \sum_{i=1}^{n} o_{i,j} \right), \text{ where } j = 1, 2, \dots m \tag{4.2}$$

$$\mathbf{s}(j) = \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(o_{i,j} - \mathbf{m}(j)\right)^2} \right), \text{ where } j = 1, 2, \dots m \tag{4.3}$$

where $\mathbf{m}, \mathbf{s} \in \mathbb{R}^{1 \times m}$.

Afterwards, we acquire the Z-score matrix $\mathbf{Z}$ from the raw opinion score matrix $\mathbf{O}$ as:

$$\mathbf{Z} = (\mathbf{O} - \mathbf{Im})./\mathbf{Is} \tag{4.4}$$

where $\mathbf{I} = [1, 1, \dots 1]^T$, $\mathbf{I} \in \mathbb{R}^{n \times 1}$ and ./ is element-wise division.

Each element $z_{i,j}$ in matrix $\mathbf{Z}$ represents the number of standard deviations by which the opinion score $o_{i,j}$ is away from the m_j. The following analyses are mainly based on the Z-score matrix $\mathbf{Z}$.

Subject bias and inconsistency

Let $\mathbf{B} \in \mathbb{R}^{1 \times n}$ and $\mathbf{C} \in \mathbb{R}^{1 \times n}$ the vector of bias and inconsistency of n subjects respectively. Bias and inconsistency for subject i is calculated with the mean and standard deviation of the Z-score for subject i across all stimulus:

$$\mathbf{B}(i) = \left(\frac{1}{m} \sum_{j=1}^{m} z_{i,j} \right), \text{ where } i = 1, 2, \dots n \tag{4.5}$$

$$\mathbf{C}(i) = \left(\sqrt{\frac{1}{m} \sum_{j=1}^{m} \left(z_{i,j} - \mathbf{B}(i)\right)^2} \right), \text{ where } i = 1, 2, \dots n \tag{4.6}$$

The key distinction between the estimation of subject bias in ZREC and P913 is that ZREC describes the subject bias in the standard deviation range of each stimulus, while P913 describes it in the opinion score range. By modeling subject bias in the standard deviation range of individual stimuli, ZREC takes stimulus ambiguity into account.

Content ambiguity

It is important to clarify the difference between stimuli ambiguity and content ambiguity. In QoE datasets, multiple stimuli (*i.e.,* PVS) can be generated from a unique source content (*i.e.,* SRC). We define the stimulus ambiguity for j as the standard deviation of

subjects' opinion score (Eq.(4.3)). Consequently, content ambiguity is defined as the mean ambiguity of all stimuli that belong to a particular content l:

$$\mathbf{A}(l) = \left(\frac{1}{h} \sum_{j \in \mathbf{g}} \mathbf{s}(j) \right), \text{ where } l = 1, 2, \dots t \tag{4.7}$$

h is the number of stimulus for content l, $\mathbf{g}$ the list of stimulus index of content l, t is the total number of contents in the entire datasets, $\mathbf{A} \in \mathbb{R}^{1 \times t}$.

Mean opinion score recovery

We first remove the bias of each subject for each stimuli from the original annotation $\mathbf{O}$. The unbiased opinion score matrix $\mathbf{U} \in \mathbb{R}^{n \times m}$ is calculated with:

$$\mathbf{U} = \mathbf{O} - \mathbf{B}^{\mathbf{T}}\mathbf{s} \tag{4.8}$$

$u_{i,j}$ is the element of $\mathbf{U}$, which is the opinion score of subject i for stimuli j after the removal of the bias of subject i. In Eq.(4.8), we multiply the bias of each observers $\mathbf{B}^{\mathbf{T}} \in \mathbb{R}^{n \times 1}$ with the standard deviation $\mathbf{s} \in \mathbb{R}^{1 \times m}$ of different subjects' opinion score for every stimuli in Eq.(4.3) in order to re-scale the Z-score to the original opinion score range.

To calculate the recovered MOS of stimuli j, denoted $\mathbf{R}(j)$, we employ a weighting scheme that takes into account the inconsistency of the opinion scores provided by different subjects. Specifically, instead of simply averaging the unbiased scores u_j across all subjects, we use a weighted average of u_j, where the weight assigned to each score is inversely proportional to the square of subject's inconsistency. This means that subjects with higher inconsistency are given less weight, and their opinion scores have less influence on the final MOS calculation.

$$\mathbf{R}(j) = \left(\frac{\sum_{i=1}^{n} \mathbf{C}(i)^{-2} u_{i,j}}{\sum_{i=1}^{n} \mathbf{C}(i)^{-2}} \right), \text{ where } j = 1, 2, \dots m \tag{4.9}$$

Similar with the recovered MOS, weighted standard deviation is calculated as:

$$\boldsymbol{\sigma}_w(j) = \left(\sqrt{\frac{n}{n-1} \times \frac{\sum_{i=1}^{n} \mathbf{C}(i)^{-2} (u_{i,j} - \mathbf{R}(j))^2}{\sum_{i=1}^{n} \mathbf{C}(i)^{-2}}} \right) \tag{4.10}$$

Where $j = 1, 2, ...m$. The factor of $n/(n-1)$ is intended to account for the number of degrees of freedom, thus giving us an unbiased estimation of the standard deviation of the population [19]. The 95% CI is thus computed with:

$$\mathbf{CI}(j) = \mathbf{R}(j) \pm 1.96 \frac{\sigma_w(j)}{\sqrt{n}} \tag{4.11}$$

P_{th} percentile opinion score recovery

Algorithm 2 Calculate Q_p, weighted p_{th} percentile opinion scores

Require: unbiased subject opinions matrix, $U_{n,m}$
Require: subject inconsistencies, C_n
Require: percentile to be calculated, p
number of subjects $= n$, number of stimuli $= m$
total weight of the population, $w = \text{sum}(C_n^{-2})$
percentile weight, $w_p = w \times p/100$
initialize Q_p, a zero vector (with size=m) to store p_{th} percentile opinion score for each stimuli
for each stimuli j in m **do**
$U_n \leftarrow$ get subject opinions for stimuli j from $U_{n,m}$
$U_{n-sorted} \leftarrow$ sort U_n in ascending order
$w_{n-sorted} \leftarrow$ sort C_n^{-2} with same indices as $U_{n-sorted}$
initialize current weight, $w_c = 0$
initialize current subject index, $i = 0$
while $w_c < w_p$ **do**
$Q_p(j) \leftarrow$ get current subject (i) opinion from $U_{n-sorted}$ and set it as the p_{th} percentile score
$w \leftarrow$ get current subject (i) weight from $w_{n-sorted}$
$w_c \leftarrow w_c + w$
$i \leftarrow i + 1$
end while
end forreturn Q_p

Some subjective/objective studies are not interested in mean of opinion scores but the percentile of opinion scores. For JND and SUR studies [154, 151, 152, 86, 34, 135], 75%SUR is commonly used to train and evaluate objective metric. It can be easily proved that for a given stimuli j:

$$q\%\text{SUR}(o_i) = (1-q) \text{ - } th \text{ percentile}(o_i), \tag{4.12}$$

75%SUR is in fact 25_{th} percentile. Therefore, we provide a weighted percentile approach where subject bias and inconsistency is taken into account. Algorithm 2 depicts the process to calculate weighted percentiles Q_p of an unbiased opinion score matrix $U_{n,m}$ of n subjects and m stimuli for a given percentile p in range [0, 100].

4.2.4 Experiment Results

MOS recovery and confidence intervals

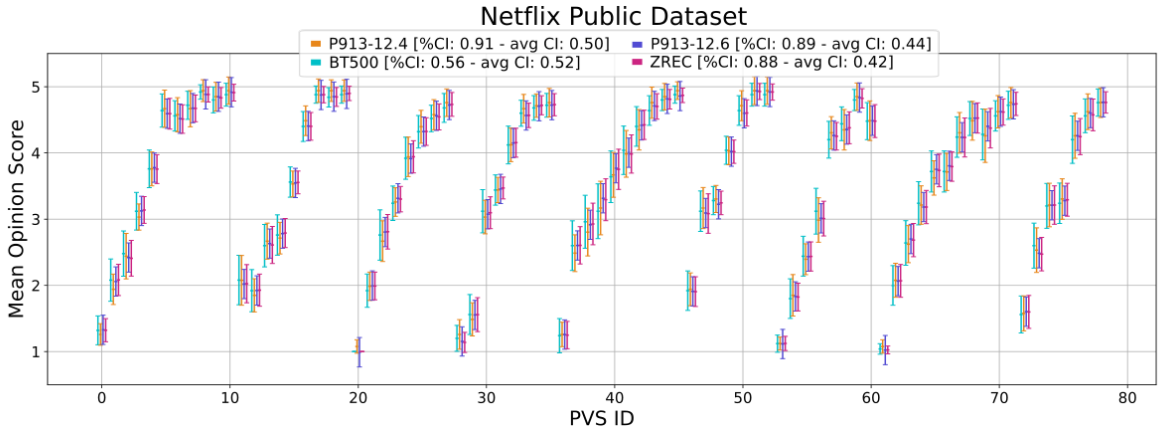


Figure 4.1 – Recovered MOS and their CI for four different methods on the Netflix Public dataset with all the observers. The legend shows the average confidence interval for each method across 79 stimuli. %CI represents the percentage of recovered MOS values that fall within the confidence interval range for the 79 stimuli, based on 1000 bootstrapping iterations. In each iteration, only half of the observers are used to recover the MOS.

Figure 4.1 depicts the MOS values and corresponding confidence intervals recovered on the Netflix Public dataset. The results indicate that P910 and ZREC offer more precise confidence intervals with values of 0.44 and 0.42, respectively.

Table 4.2 depicts the average 95% CI of the recovered MOS on the Netflix Public and AMZ-HD-VJND datasets with all subjects. To evaluate the comparative reliability of the confidence intervals generated by each method, we performed a bootstrapping analysis comprising 1000 iterations. In each iteration, we randomly selected half of the subjects and recovered the MOS with each model. Results shows that ZREC and P910 exhibit the lowest average CI values while maintaining a relatively high CI% level. Despite having a higher CI% value, P913 displays a significantly larger average CI compared to ZREC and P910.

Table 4.2 – Analysis of the CI of the recovered MOS for four different methods on the Netflix Public and the AMZ-HD-VJND datasets. Avg CI represents the length of the CI for each method on each dataset with all subjects included in MOS recovery. CI% represents the percentage of recovered MOS values that fall within the confidence interval range based on 1000 bootstrapping iterations. In each iteration, only half of the subjects in each dataset are used to recover the MOS.

	NETFLIX		AMZ-HD-VJND	
	Avg CI	CI%	Avg CI	CI%
BT500	0.5153	0.5645	1.2612	0.9285
P913	0.4986	0.9102	1.1254	0.8671
P910	0.4420	0.8885	1.0217	0.8805
ZREC	0.4172	0.8783	0.9813	0.8554

Estimated parameters

Table 4.3 – Pearson linear correlation coefficient (PLCC) between the estimated parameters of subject inconsistency, subject bias, and content ambiguity across various models.

	Subject Inconsistency	Subject Bias	Content Ambiguity
Model Pair	NETFLIX		
MLE - ZREC	0.9282	0.9952	0.9663
MLE - P910	0.9669	0.9964	-
P910 - ZREC	0.9372	0.9965	-
P913 - MLE	-	0.9992	-
P913 - P910	-	0.9999	-
P913 - ZREC	-	0.9965	-
Model Pair	AMZ-HD-VJND		
P910 - ZREC	0.9603	0.9994	-
P913 - P910	-	0.9999	-
P913 - ZREC	-	0.9994	-

In this section, we analyze the correlation between the subject bias, inconsistency and content ambiguity across the tested models. As summarized in the Table 4.1, BT500 does not estimate any of the parameters and thus excluded from the correlation analysis. Moreover, P913 cannot estimate subject inconsistency and content ambiguity while P910 cannot estimate content ambiguity. In addition, MLE fails to converge to a solution for AMZ-HD-VJND dataset.

Table 4.3 depicts the PLCC values between the indicated model pairs in each row. The results indicate that the tested models are well correlated in terms of subject bias. On the other hand, estimated subject inconsistencies show slight differences between models.

Finally, MLE and ZREC shows relatively high correlations in terms of content ambiguity. Despite the lower correlations for subject inconsistencies, ZREC estimations are in line with the standards. It is impossible to know which model estimations are closer to the ground truth, however the analysis showcases the relative reliability of the approach.

Impact of percentile opinion score recovery on the accuracy of SUR prediction models

Table 4.4 – The mean and variance of absolute errors on 75%SUR prediction with SUR prediction model [154] on AMZ-HD-VJND dataset without any recovery and with ZREC and P910 POS recovery.

|Δ75%SUR|	Without Recovery [154]	P910 POS Recovery	ZREC POS Recovery
Mean Error	0.7489	0.7175	**0.6883**
Error Variance	0.9224	0.7198	**0.6989**

Previous work [118] has shown that training objective quality models on cleaned data can improve the prediction performance. In this work, we compared the performance of the 75%SUR prediction model [154] trained on 75%SUR from original datasets without recovery and 75%SUR (25_{th} percentile) recovered by ZREC and P910 respectively. Because P910 only provide MOS recovery but not percentile recovery, we use Algorithm 2 with the subject bias and inconsistency of P910 as input. The mean and variance of absolute error of 75%SUR for different training data are shown in Table 4.4. It can be observed that the 75%SUR prediction model trained both on ZREC and P910 improved the prediction, in which ZREC get a smaller prediction error than P910.

4.2.5 Conclusion

We introduced ZREC to estimate subject bias, inconsistency and content ambiguity, all of which are fundamental for QoE studies. Using these parameters, ZREC can recover the MOS and the POS whichever is more suitable for the QoE use-case in question. Our findings indicate that ZREC can produce slightly tighter CIs for MOS recovery on two datasets compared to the current state of the art models, albeit with a minor reduction in accuracy. A tighter CI allows to reduce the required number of subjects in the subjective study without sacrificing from the accuracy and resolving power. Furthermore, the results of our experiments on the SUR prediction use-case demonstrate that ZREC can improve

the performance of objective quality metrics by providing a more reliable ground truth with 25_{th} POS recovery.

4.3 Uncertainty analyses of SUR

In this section, we emphasize the importance of uncertainty estimation for subjective SUR, a factor often overlooked in previous works (Section 4.3.1). We then introduce a method to estimate the uncertainty of p%SUR (Section 4.3.2), which represents a single point on the SUR curve, followed by estimating the uncertainty for the entire SUR curve (Section 4.3.3).

For p%SUR, our CI estimation method does not rely on any distribution assumption for the individual Just Noticeable Difference (JND). However, for SUR curve CI estimation, distribution assumption becomes necessary. Additionally, we validate each mathematical CI estimation method using the bootstrapping method.

4.3.1 Motivation

Several studies have shown that depending only on the Mean Opinion Score (MOS) isn't enough because it overlooks the diversity in subjective ratings [47]. Additionally, it is important for VQM to consider the Confidence Interval (CI) of subjective data [119, 28, 25]. Ignoring the CI/uncertainty can lead to training models based on data that are not statistically significant, resulting in an inaccurate understanding of the correct behavior [70].

Similarly, the Confidence Interval (CI) of q%SUR, which represents the P_{th} percentile of the individual Just Noticeable Difference (JND) score (see Eq.(4.12)), is also important for further analysis of subjective data and the development of objective metrics based on SUR.

As shown in Figure 4.2, when revisiting the original annotations in VideoSet, we discovered instances where certain SRCs depicted nearly identical scenes, yet their respective 75%SUR values exhibited considerable disparity. As depicted in Figure 4.2, we present sample frames from SRC#76 and #79, both featuring nearly identical video content, along with the distributions of original VW-JND annotations provided by the individuals. Notably, the 75%SUR QP values for these two SRCs are 33 and 30, respectively, indicating a significant difference. However, when performing the ANOVA [39] analysis on the two

distributions, no statistically significant differences emerged. Furthermore, when examining the 95% confidence interval (95%CI) ranges for the 75%SUR values, they appear relatively close, despite the significant disparity in the 75%SUR values themselves.

It is worth highlighting that previous works [137, 152, 135] have employed the 75%SUR as ground truth for training their models, aiming to predict two distinct values for what are essentially the same video contents, which can cause ambiguity for model training. Therefore, it becomes imperative to analyze the uncertainty associated with the SUR derived from subjective tests.

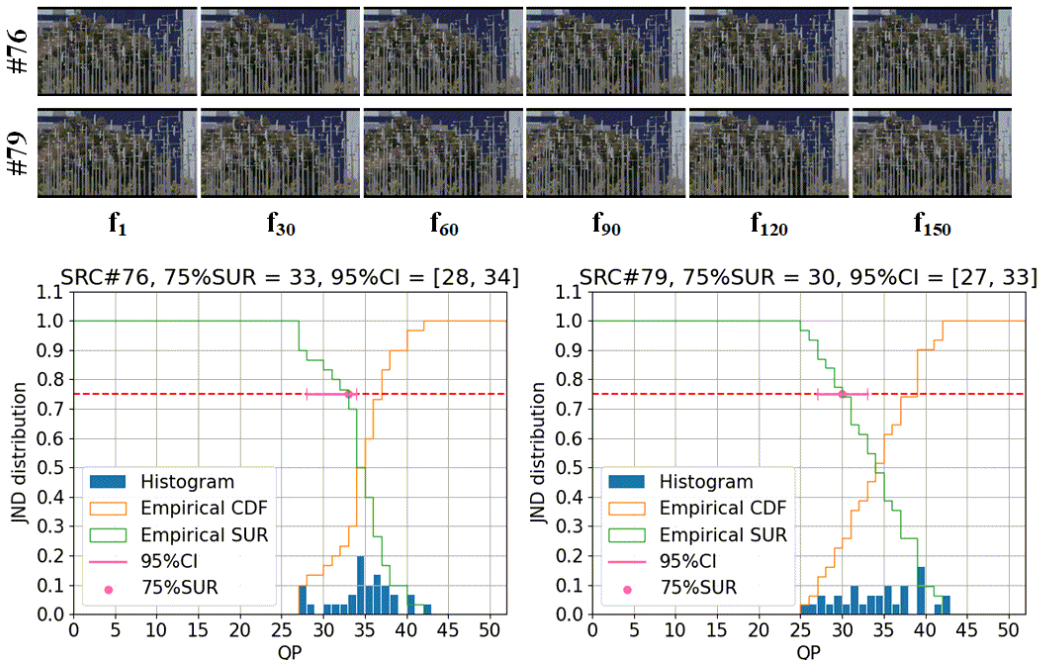


Figure 4.2 – The sampled frames of SRC#76 and #79 and the corresponding VW-JND distribution in VideoSet [138].

4.3.2 Uncertainty estimation of p%SUR$_{\text{emp}}$

Mathematical CI estimation

In this section, we apply the definition of SUR for various proxies as outlined in Section 3.2.2. We can determine p%SUR$_{\text{emp}}$ for a specific video content using individual VW-JND annotations collected from a sampled population through subjective test. However, if we were to replicate the same test with a different group of subjects, would we obtain the same p%SUR$_{\text{emp}}$ results?

Figure 4.2 has shown that the 75%SUR$_{\text{emp}}$ for almost same contents can be very different. Therefore, assessing the uncertainty of the p%SUR$_{\text{emp}}$ data obtained from the collected datasets is very important.

Using statistical theory, we can estimate the true p%SUR of the entire population based on the p%SUR$_{\text{emp}}$ obtained from a sample of N subjects. If we assume that the true p%SUR is equal to s, and we randomly select one subject from the population with their VW-JND denoted as j_n^m, we can calculate the probability of j_n^m being less than s using Eq.(4.13), in accordance with the definition of the p%SUR in Section 3.2.2. As a reminder, in case 1, where quality decreases with an increase in the proxy (e.g., using QP as the proxy as shown in Figure 4.2), the empirical SUR corresponds to the complementary empirical CDF. In contrast, in case 2, where quality increases with the proxy (such as VMAF), the empirical SUR is the empirical CDF itself.

$$Pr(j_n^m \leqslant s) = \begin{cases} (1-p)\%, \text{for case 1,} \\ p\%, \text{for case 2.} \end{cases} \tag{4.13}$$

Taking case 2 as an example, we define the random variable A as equal to 1 (event success) when $j_n^m \leqslant s$ and 0 (event failure) when $j_n^m > s$. Consequently, the random variable A conforms to a Bernoulli distribution [18], as presented in Table 4.5.

Table 4.5 – The random variable A follows a Bernoulli distribution (this table serves as an example for case 2)

Event	A	Probability
$j_n^m \leqslant s$	1 (success)	$p\%$
$j_n^m > s$	0 (fail)	$(1-p)\%$

A subjective test involving N subjects can be understood as N times independently sampling the population. The count of event successes, denoted as X, conforms to a binomial distribution [36]:

$$X \sim B(N, p\%). \tag{4.14}$$

The PMF of X can be obtained by:

$$f(x, N, p\%) = \Pr(X = x) = C_N^x p\%^x (1-p\%)^{N-x} \tag{4.15}$$

Where $C_N^x = \frac{N!}{x!(N-x)!}$ and $x \in [0, N]$. Figure 4.3 shows the PMF of the binomial distribution with parameters $N = 34$ and $p = 75$. When the count of event successes is

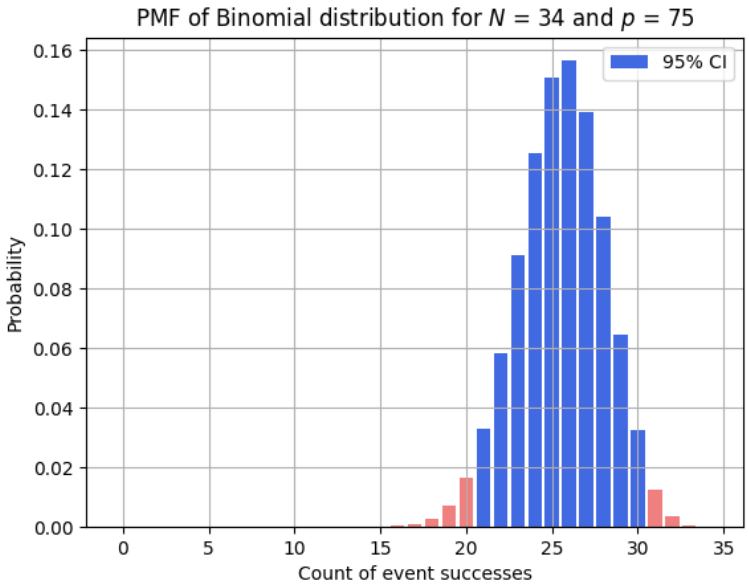


Figure 4.3 – PMF of binomial distribution for N=34, p=75, and the 95%CI of 75%SUR

26, the probability is calculated as 0.1564. This indicates that if we were to conduct a subjective test with 34 subjects, there is a 15.64% probability that 26 of these subjects would have VW-JND values smaller than or equal to s. If we can determine the lower and upper bounds, denoted as l and u, respectively, such that the cumulative probability between them encompasses approximately 95%, we can confidently assert that there is a 95% probability that the number of subjects with $j_n^m \leqslant s$ falls within the interval $[l, u]$.

We adapted the Near-symmetric Algorithm [41] to derive l and u for the desired CI as described in Algorithm 3. Once l and u are determined, we arrange the values of J^m in ascending order. Subsequently, the CI range for $p\%\text{SUR}_{\text{emp}}$ is between $CI_l = J^m_{\text{ordered}}[l]$ and $CI_u = J^m_{\text{ordered}}[u]$, where J^m_{ordered} represents the ordered values of J^m. The 95%CI range can be interpreted as follows: if we were to replicate the subjective test multiple times, there is a 95% probability that the $p\%\text{SUR}_{\text{emp}}$ falls within this range.

Bootstraping CI

After computing the 95%CI ranges as presented previously, we perform bootstrapping on the original annotations to compare with the CI estimation. For each bootstrap sample, we computed $p\%\text{SUR}_{\text{emp}}$ and calculated the percentage of $p\%\text{SUR}_{\text{emp}}$ values that fell within the estimated 95%CI, denoted as Avg CI. We performed 1,000,000 bootstrapping

Algorithm 3 Get lower bound and upper bound of target CI

Input: $target\_CI$, **binomial** (index from 0 to N)

$max\_index \leftarrow argmax(\textbf{binomial})$
$real\_CI\_list = []$; $l\_list = []$; $u\_list = []$
$l, u \leftarrow max\_index$; $real\_CI \leftarrow \textbf{binomial}[max\_index]$
while $real\_CI < target\_CI$ **do**
    $left \leftarrow \textbf{binomial}[l-1]$
    $right \leftarrow \textbf{binomial}[l+1]$
    if $left \leqslant right$ **then**
        $real\_CI += left$; $l = l - 1$
    else
        $real\_CI += right$; $u = u + 1$
    end if
    $real\_CI\_list.append(real\_CI)$; $l\_list.append(l)$; $u\_list.append(u)$
end while
if $real\_CI\_list[-1] - target\_CI > target\_CI - real\_CI\_list[-2]$ **then**
    $real\_CI = real\_CI\_list[-2]$, $l = l\_list[-2]$, $u = u\_list[-2]$
else
    $real\_CI = real\_CI\_list[-1]$, $l = l\_list[-1]$, $u = u\_list[-1]$
end if
return l, u, $real\_CI$

iterations, each with sample sizes of 0.25, 0.5, and 0.75 of the original annotations. Table 4.6 shows the Avg CI values for 95%CI estimation on 220 video contents of VideoSet in 1080p for 1st JND.

Table 4.6 – Avg CI with 1,000,000 bootstrapping iteration with different sample sizes

Sample size	0.25	0.5	0.75
Avg CI	0.8331	0.9790	0.9998

In VideoSet, each SRC is annotated by 25 to 34 subjects. Consequently, when the sample size is reduced to 0.25, we observe a decrease in Avg CI. However, on average, the bootstrapped CI closely aligns with the mathematically based CI estimation presented in the previous section, confirming the validity of our proposed mathematical CI estimation method.

4.3.3 Uncertainty estimation of SUR curve

In the preceding section, we calculated the uncertainty associated with the $p\%$ SUR_{emp}, which represents just one point of the SUR curve. In this section, we delve into a comprehensive analysis of the uncertainty across the entire SUR curve. This is accomplished through the application of Maximum Likelihood Estimation (MLE) to fit an analytical curve.

MLE estimation

The VW-JND of each video content clip m can be seen as a random variable J^m. The annotations from the JND subjective test by N observers can be seen as N independent and identically distributed (*i.i.d.*) samples of J^m. From the observed values, *i.e.,* the vector of subjects' annotations in Eq. (3.1), we can estimate the distribution of J^m.

Previous works [138, 137, 135, 152, 151, 63] have assumed that J^m follows a Gaussian distribution. We can use Maximum Likelihood Estimation (MLE) to estimate the parameters of this distribution following the Gaussian assumption. The probability density function of the Gaussian distribution for video content clip m is given by:

$$f^m(j|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{j-\mu}{\sigma}\right)^2}, \tag{4.16}$$

where μ and σ are two parameters of Gaussian distribution. The likelihood function of clip m:

$$L^m(\mu,\sigma^2|\mathbf{j}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2}, \tag{4.17}$$

where j_i^m is the VW-JND value of the i-th observer obtained from the subjective test for clip m. The log-likelihood function of clip m:

$$\ell^m(\mu,\sigma^2) = \log(L^m(\mu,\sigma^2|\mathbf{j})) \tag{4.18}$$

The gradient vector of the log-likelihood function of clip m:

$$\mathbf{u}(\theta) = \frac{\partial \ell^m(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \mu}\ell^m(\mu,\sigma^2) \\ \frac{\partial}{\partial \sigma^2}\ell^m(\mu,\sigma^2) \end{pmatrix}, \tag{4.19}$$

where $\mathbf{u}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times 1}$, p is the number of the parameters. For Gaussian distribution, $p = 2$.

The optimal $\hat{\mu}^m$ and $\hat{\sigma}^m$ are typically found by setting the gradient of the log-likelihood

to zero, which is a necessary condition for a maximum. However, it is not always sufficient, because any point where the gradient is zero could be a local maximum, a local minimum, or a saddle point.

$$\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0} \tag{4.20}$$

This procedure can be generalized to other distributions than Gaussian, such as Logistic, Weibull, Gumbel, Rayleigh, etc.

Goodness-of-fit

However, when we observer the raw distribution of the VideoSet [138], we noticed that the Gaussian distribution may not be the most suitable to model the data, as shown in Figure 3.2 in Chapter 3. Therefore, we explore several alternative distributions and conduct goodness-of-fit tests to identify the most appropriate one. The experimental results of the goodness-of-fit tests are presented in Table 4.7. The results reveal that the Weibull distribution yields the largest log-likelihood, indicating its superiority in modeling the VW-JND data of a group of observers.

Table 4.7 – Goodness of fit: different distributions and the resuls of log-likelihood

Distribution	Nb of para	Nb of reject	Log-likelihood
Gaussian	2	2	-93.8587
Logistic	2	0	-94.1238
Weibull	2	0	**-93.6152**
Gamma	2	13	-96.2992
Gumbel	2	3	-97.0893
Rayleigh	1	218	-117.0835
Cauchy	2	1	-98.9519
Student-t	3	0	-93.6944

Mathematical CI estimation of MLE

The parameters of the distribution of individual JND can be estimated by the samples of individual JND from subjective test using MLE. However, how certain is the estimation? If we repeat the same subjective test with another group of observers, the estimated parameters will probably change. Therefore it is important to estimate the Confidence Interval (CI) of the estimated parameters of MLE.

Bradley [22] outlined a method for estimating confidence intervals (CIs) for Maximum Likelihood Estimators (MLEs). We adopt Bradley's approach to determine the CIs for the MLEs.

Initially, we compute the *observed* Fisher's Information matrix, denoted as $\boldsymbol{I}(\theta)$ defined as:

$$\boldsymbol{I}(\theta) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta), 1 \leq i, j \leq p \tag{4.21}$$

Here p represents the number of parameters, typically $p = 2$ for Gaussian distributions. The observed Fisher's Information matrix is the negative of the Hessian matrix of the log-likelihood function $\ell(\theta)$.

The *expected* Fisher's Information matrix $\boldsymbol{\mathcal{I}}(\theta)$ is the expectation of the observed Fisher's Information matrix:

$$\boldsymbol{\mathcal{I}}(\theta) = -E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta)\right), 1 \leq i, j \leq p \tag{4.22}$$

Similar to the Central Limit Theorem (CLT), when the sample size N is large and the J^m are independent and identically distributed (i.i.d.) random variables, the distribution of estimated parameters by MLE approaches normality asymptotically. Here, the mean of this distribution equals the true parameter value, while the variance-covariance matrix mirrors the inverse of the expected Fisher's information matrix, denoted as $\boldsymbol{\mathcal{I}}(\theta)^{-1}$. Within this matrix, the diagonal elements correspond to the variances of the estimated parameters, and the off-diagonal elements represent their covariances.

Mathematically, this can be expressed as:

$$\theta \sim \mathcal{N}(\theta, Diag\left(\boldsymbol{\mathcal{I}}(\theta)^{-1}\right)) \tag{4.23}$$

Since the true parameters are typically unknown, we rely on MLE estimated parameters to compute confidence intervals. The 95% confidence interval of the parameters is calculated as follows:

$$\hat{\theta} \pm z_{value} \times \sqrt{Diag\left(\boldsymbol{\mathcal{I}}(\hat{\theta})^{-1}\right)} \tag{4.24}$$

For a detailed mathematical demonstration of the confidence interval estimation using MLE, please refer to Annex F.

After estimating the CIs for each parameter, we can determine the CI of the SUR curve. We can plot the SUR curves with the lower and upper bounds of the CIs to visualize the

uncertainty of the SUR curve, as shown in Figure 4.4.

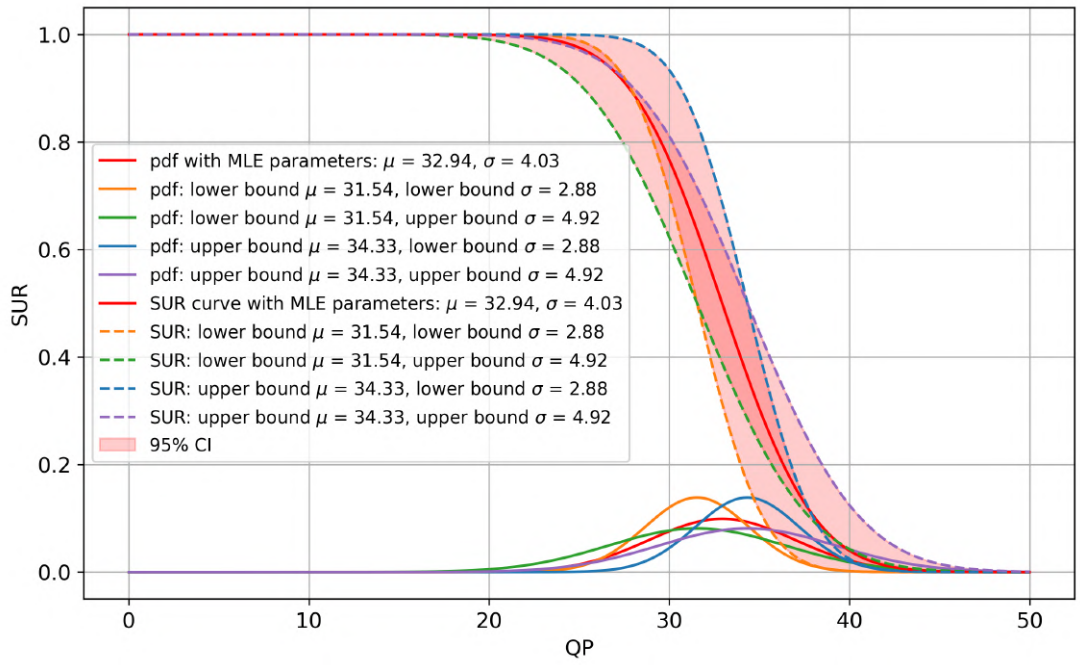


Figure 4.4 – MLE estimated SUR curve and 95%CI for SRC12 in VideoSet [138] 1080p with Gaussian assumption.

The solid lines represent the estimated Probability Density Functions (PDF), while the dashed lines depict the SUR curves. The shaded areas indicate the 95% confidence intervals (CIs) of the SUR curve. The red curves depict the MLE estimation without the CI, whereas the other four colors represent combinations of the lower and upper bounds of the estimated parameters μ and σ with Gaussian assumption.

The CI of the SUR curve indicates the uncertainty inherent in the subjective test results. For instance, taking SRC12 of VideoSet (Figure 4.4) as an example, at the 50%SUR level, there is a 95% probability that the QP value ranges from 31 to 34. Conversely, for QP = 30, there is a 95% chance that the SUR value falls between 0.61 and 0.93. This implies that between 61% and 93% of observers may not perceive any difference between QP 0 and QP 30 for SRC12.

Considering this uncertainty is crucial when utilizing the SUR curve for further analysis or model training.

Bootstraping CI Validation

Similar to the CI estimation of p%SUR in Section 4.3.2, we employ bootstrapping to validate the mathematical CI estimation of the MLE parameters. We execute 1,000 bootstrapping iterations, each with sample sizes of 0.25, 0.5, and 0.75 of the original annotations. MLE is performed on each subset of the bootstrapped data, and the percentage of MLE parameters falling within the mathematically estimated 95% CI, as outlined in Section 4.3.3, is computed and denoted as Avg CI. The results of the Avg CI for the 95% CI of VideoSet 1080p are presented in Table 4.8.

Table 4.8 – Avg CI with 1,000 bootstrapping iteration with different sample sizes

Sample size		0.25	0.5	0.75
Avg CI	μ	0.6812	0.8380	0.9129
	σ	0.6811	0.8546	0.9265

Similar with the CI estimation of p%SUR$_{\text{emp}}$, the Avg CI increases as the sample size increases. When sample size is 0.75, the Avg CI of μ and σ are close to around 91.29% and 92.65%, respectively. This indicates that the mathematically estimated 95% CI of the MLE parameters is reliable and can be used for further analysis.

4.4 Longitudinal study of Subjective Data

One of the unique features of our AtHome subjective test pipeline is that participants conduct subjective tests over the long term (e.g., 30 minutes per day for 20 days), in contrast to traditional InLab tests or crowdsourcing tests where participants conduct the test in a single session (e.g., 30 minutes). This feature allows us to study the longitudinal effects of subjective data. In this section, we will examine the longitudinal effects of subjective data by analyzing the subjects' behavior.

4.4.1 Test campaign management

Different from the traditional subjective test conducted in a single session over a short time, the AtHome subjective test pipeline allows participants to conduct the test over the long term. That is to say, participants can choose to connect to the server and conduct the test at any time during several days. To prevent participant fatigue, we've implemented

time limits on the application. For instance, participants cannot conduct the test for more than 45 minutes per day, and they are asked to finish the test within a deadline.

For example, collecting a JND dataset with 180 video contents will take around 9 hours per participant. We require participants to complete the entire test campaign within 3 weeks. One participant can choose to connect to the server and conduct the test for 30 minutes per day for 18 days to meet our requirements.

As illustrated in Figure 4.5, participants will receive email notifications announcing the beginning of the next test campaign, detailing the total time needed to finish the test, the daily time limits, and the campaign deadline.

Participants who are unable to finish the test need to return the TV and other materials to us. We will then provide the TV to another participant to continue the test. However, this may result in additional delays to the test campaign.

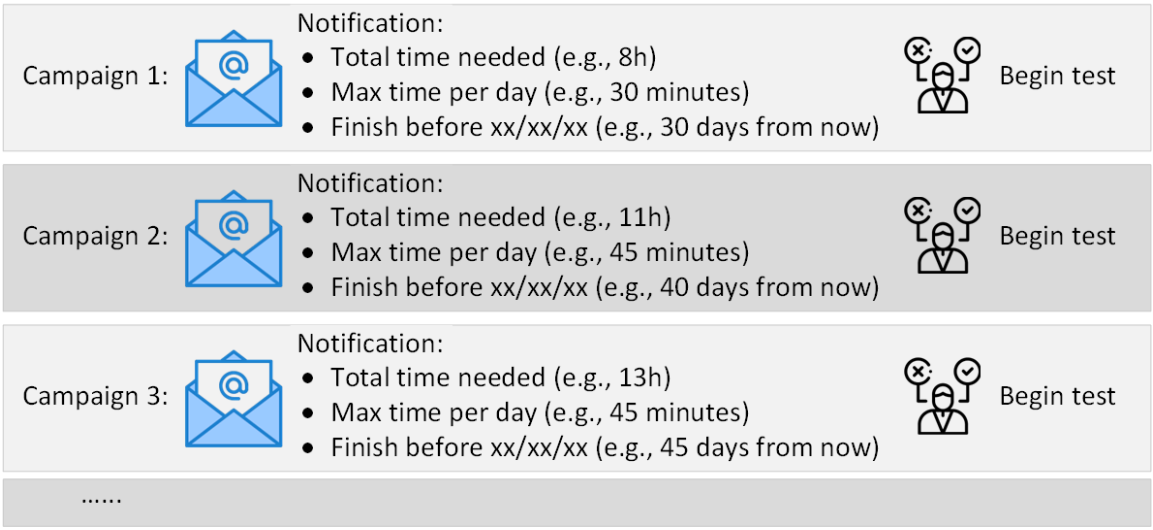


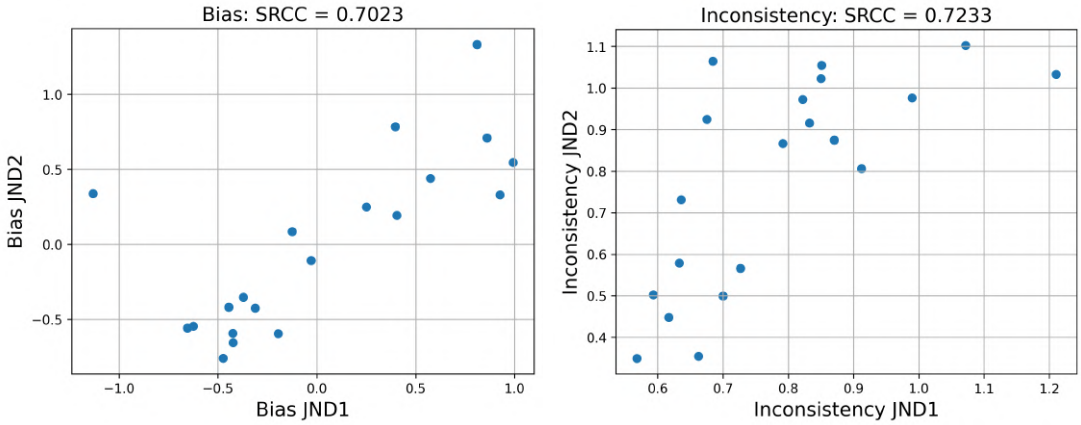
Figure 4.5 – Demostration of different test campaigns of the AtHome subjective test pipeline.

4.4.2 Observer behavior analysis

The subjective test data collected following the test campaign management described above enables us to conduct longitudinal analysis of the observers' behavior. In this section, we perform **cross campaign** analysis, where we compare observer behavior across different test campaigns, and **intra campaign** analysis, where we compare observer behavior across different days within the same test campaign. Observer behavior is characterized by the bias and inconsistency of the observers, quantified by ZREC (see Section 4.2).

Cross campaign analysis

The same group of observers conducts the two JND test campaigns (namely JND1 and JND2). We compute the bias and inconsistency of each observer in JND1 and JND2 using ZREC. Note that JND1 corresponds to the first JND search, while JND2 corresponds to the second JND search. For more details, please refer to Figure 3.1. The bias and inconsistency of the observers in JND1 and JND2 are shown in Figure 4.6. The results indicate that the bias and inconsistency of the observers in JND1 and JND2 are correlated (with Spearman rank correlation coefficients of 0.7023 and 0.7233 respectively), even though there is a significant time gap between the two test campaigns (there is a gap of one month between JND1 and JND2). This suggests that the bias and inconsistency of the observers are relatively stable over time.



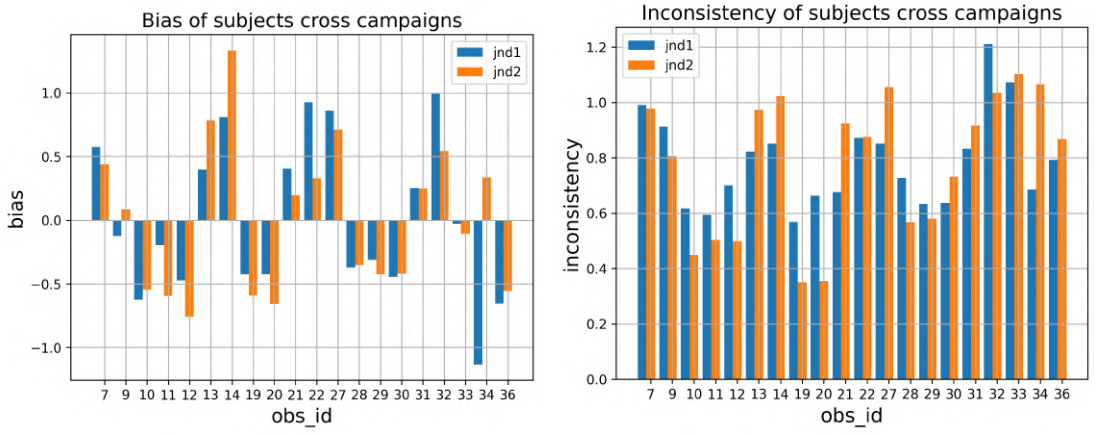
(a) (b)

Figure 4.6 – Correlation analyses of observer bias and inconsistency of test campaign JND1 and JND2.

From Figure 4.7a, we can observe that only one observer (ID 34) changed his bias

drastically from JND1 to JND2. For some observers, the bias is reduced from JND1 to JND2, while for others, it increases.

The inconsistency of the observers in JND1 and JND2 is shown in Figure 4.7b. The inconsistency of the observers in JND1 and JND2 is slightly more highly correlated than the bias, indicating that the time gap has less impact on the inconsistency of the observers than on the bias. Similar to bias, there is no clear trend in the inconsistency of the observers from JND1 to JND2. This indicates that the time gap does not have a significant effect on reducing or increasing the inconsistency of the observers between two test campaigns.



(a) (b)

Figure 4.7 – Bias and inconsistency of all observers in different test campaigns.

Intra campaign analysis

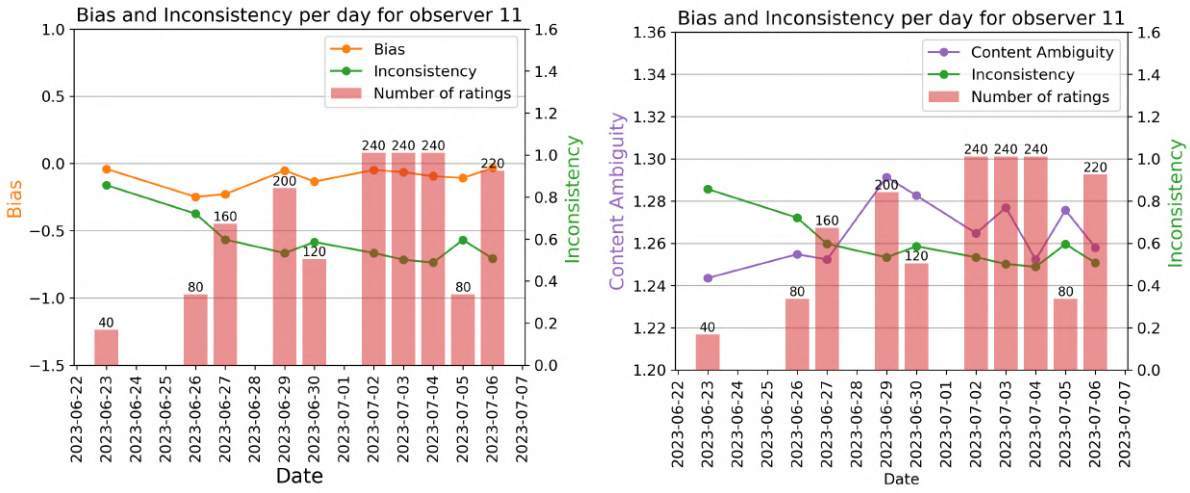
Each test campaign lasts for several days (e.g., 20 days). We can also analyze the observer behavior within each test campaign, which is referred to as intra-campaign analysis. We compute the bias and inconsistency of each observer for each day of the test campaign using an adaptation of ZREC.

Instead of computing the bias and inconsistency of each observer i across the entire test campaign as in Eq.(4.5) and Eq.(4.6), we compute the bias and inconsistency of each observer i for each day d of the test campaign. The bias and inconsistency of observer i on day d are denoted as $\mathbf{B}_d(i)$ and $\mathbf{C}_d(i)$ respectively. The bias and inconsistency of observer i on day d are computed as follows:

$$\mathbf{B}_d(i) = \left(\frac{1}{D} \sum_{j \in d} z_{i,j} \right), \text{ where } i = 1, 2, \ldots n \tag{4.25}$$

$$\mathbf{C}_d(i) = \left(\sqrt{\frac{1}{D} \sum_{j \in d} \left(z_{i,j} - \mathbf{B}_d(i)\right)^2} \right), \text{ where } i = 1, 2, \ldots n \tag{4.26}$$

Where D is the number of stimuli done by observer i on day d. $z_{i,j}$ is computed from Eq.(4.4). $j \in d$ means that the stimuli j is done by observer i on day d.



(a) (b)

Figure 4.8 – Intra campaign analysis of observer 11. (a) Bias and inconsistency of observer 11 for each day of the test campaign. (b) Ambiguity and inconsistency of observer 11 for each day of the test campaign.

Figure 4.8a showcases the behavior of Observer 11. The red bars represent the number of ratings done by Observer 11 on each day. It took him 15 days to complete this test campaign. It can be observed that from July 2nd to July 4th, Observer 11 reached the maximum limit of per day test time (240 ratings take around 45 minutes). The bias of Observer 11 remains relatively stable over time, and there is a tendency for the inconsistency of Observer 11 to decrease.

In Section 4.2, we demonstrated that the observer inconsistency proposed by ZREC eliminates the influence of content ambiguity. However, here we aim to confirm whether high observer inconsistency arises because the content is overly ambiguous. Therefore, we also plot the average content ambiguity (computed using Eq. (4.7)) per day in Figure 4.8b. It can be observed that high observer inconsistency does not solely result from high content

ambiguity. For instance, the content ambiguity on June 23rd is lower than that on June 26th, yet the observer inconsistency on June 23rd is higher than that on June 26th. This observation further confirms that the observer inconsistency proposed by ZREC is independent of content ambiguity.

Similar analyses for all the other observers can be found in Annex E. It can be observed that the trend of observer bias and inconsistency over time is not consistent across all observers. In order to have a quantitative analysis of the observer behavior over time, we compute the SRCC between bias and inconsistency of each observer over time, as shown in Figure 4.9. For observers with ID 12, 18, and 28, we can see that there is a negative correlation relationship between the observer inconsistency and the date, indicating that the inconsistencies of these observers decrease as the test campaign progresses. However, this does not hold true for all observers. For instance, the observer with ID 31 has a positive correlation between the observer inconsistency and the date, indicating that the inconsistency of this observer increases as the test campaign progresses. This demonstrates that the observer behavior is not consistent across all observers.

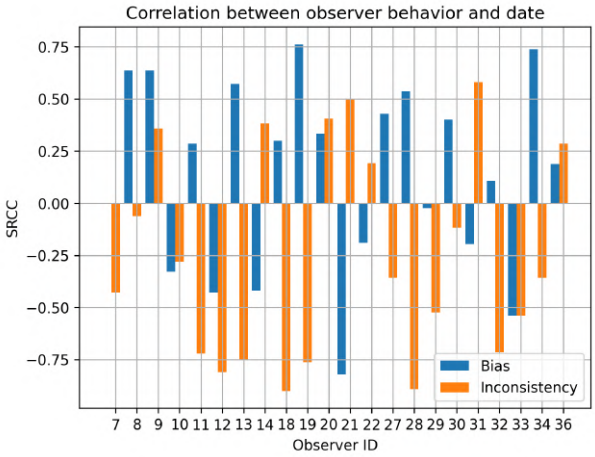


Figure 4.9 – SRCC between observer behavior and date for each observer within the same test campaign.

We compute the mean and standard deviation of the SRCC for all the observers; the results are presented in Table 4.9. The last column of the table represents the ratio between the observer inconsistency and the average content ambiguity per day. It can be observed that there is no clear trend of the observers' bias and inconsistency over time within the test campaign.

Table 4.9 – Mean and standard deviation of SRCC between observer behavior and date for all observers within the same test campaign.

Behavior	**Bias**	**Consistency**	**Ratio**
SRCC (mean±std)	0.1352 ± 0.4435	-0.2507 ± 0.4906	-0.2504 ± 0.5033

4.5 Summary

In this chapter, we first introduced our data screening method ZREC, which could be used for both MOS and POS. Besides, ZREC also provided a more straightforward way to estimate subject bias, inconsistency, and content ambiguity compared to P910. Experiment results showed that ZREC could produce slightly tighter CIs for MOS recovery compared to the state-of-the-art methods. Besides, the results of our experiments on the SUR prediction use-case demonstrated that ZREC improved the performance of objective quality metrics by providing a more reliable ground truth.

We then introduced the uncertainty estimation of the SUR curve, which was overlooked in previous works but very important for JND studies. We first proposed a mathematical method to estimate the CI for p%SUR, which was a single point giving the level of SUR value. This method didn't rely on any distribution assumption. We then estimated the uncertainty of the entire SUR curve by using the CI estimation of MLE parameters. We also validated the CI estimation of MLE parameters by bootstrapping for both methods. Further analyses based on the uncertainty of the SUR, such as examining how video quality metric prediction spread for a p%SUR value, will be conducted in the next chapter.

Finally, we conducted a longitudinal study of the subjective data collected by our At-Home subjective test pipeline. We first introduced our test campaign management, which allowed participants to conduct the test over the long term. We then analyzed the observer behavior by comparing the bias and inconsistency of the observers across different test campaigns and within the same test campaign, namely cross-campaign analysis and intra-campaign analysis. The results showed that the bias and inconsistency of the observers were relatively stable over time. The trend of observer bias and inconsistency over time was not consistent across all observers. We also computed the SRCC between bias and inconsistency of each observer over time, and the results showed that there was no clear trend of the observers' bias and inconsistency over time both for cross-campaign and intra-campaign analyses.

Chapter Contributions

— Proposed ZREC: an effective method for screening subjective data.

— Analyzed how data screening affects learning-based SUR prediction model.

— Developed mathematical methods to estimate uncertainty in empirical SUR values and analytical SUR curves.

— Conducted a longitudinal study using data collected through our AtHome subjective test pipeline.

Chapter 5

Objective Study of SUR

Overview

Contents

5.1 Introduction 83
5.2 Resolving power of VQM towards SUR 84
5.2.1 VQM 85
5.2.2 Experimental resuls 86
5.3 Prediction of SUR using VQMs as proxy 90
5.3.1 State-of-the-art methods 93
5.3.2 $\Delta VMAF_{SUR(75\%)}$ prediction pipeline using VQMs 95
5.3.3 Experimental Results 97
5.3.4 Discussion 99
5.4 Prediction of SUR using encoding parameters as proxy 100
5.4.1 Parameter-driven model 101
5.4.2 Further improvement of the prediction 107
5.5 Summary 114

Part of this chapter has been published in research papers [156, 158, 159, 160]

5.1 Introduction

We collected subjective VW-JND datasets (Chapter 3) and conducted preliminary analyses (Chapter 4), providing insights into the SUR for VW-JND. However, it is impractical to collect subjective data for every video. Therefore, developing objective models to accurately predict the SUR for VW-JND is essential. This chapter aims to address the following research questions:

— How effectively can current widely used VQMs reflect the SUR of VW-JND?

— Can we predict the SUR of VW-JND using VQMs as proxies?

— Can we predict the SUR of VW-JND using encoding parameters as proxies?

In this chapter, we delve into the objective study of the SUR by answering these research questions. We start by analyzing the ability of widely used VQMs to reflect SUR in Section 5.2. Experimental results in Section 5.2.2 reveal that VQMs exhibit high content dependency for a given SUR threshold, indicating that current VQMs are not sufficiently accurate in capturing the SUR. This finding motivates us to develop a learning-based model to predict the SUR of VW-JND using VQMs as proxies, detailed in Section 5.3. Additionally, we propose a novel parameter-driven framework (Section 5.4.1) and its improved version (Section 5.4.2) for predicting SUR using encoding parameters as proxies, presented in Section 5.4.

5.2 Resolving power of VQM towards SUR

Before developing learning-based models on the subjective datasets, we first analyzed the existing widely used VQMs. VQMs [140, 139, 80, 9] capture video quality on a continuous scale and aim to exhibit a strong correlation with human visual perception. We applied these VQMs to the VW-JND datasets to address the research question: How well can the current widely used VQMs reflect the Just Noticeable Difference (JND)? Taking VMAF as an example, can we find a threshold VMAF value at which 75% of observers cannot perceive a quality difference compared with the pristine video? We termed this the **resolving power** of VQM towards SUR. In the literature, VMAF scores of 94 and 98 are commonly used for the first JND [102, 100]. For instance, using VMAF as an example, the threshold VMAF value for the first JND can vary among different observers for the same video content. Likewise, for a given observer, different video contents may yield different threshold VMAF values for the first JND. Therefore, it is important to investigate the resolving power of VQMs towards SUR.

In this section, we investigate the resolving power of VQMs towards SUR. We compute the VQMs corresponding to a fixed SUR threshold and analyze the consistency and uncertainty of different VQMs on both publicly available VW-JND datasets and our collected VW-JND datasets across various content types.

5.2.1 VQM

We investigate the use of the following VQMs, which are the most widely used video quality metrics in practice, as proxies for SUR:

— **Peak Signal-to-Noise Ratio (PSNR)**: PSNR is a video quality metric that measures the difference between an original video signal and a distorted version of that signal. It is commonly used to assess the fidelity or visual quality of compressed or reconstructed video.

— **Structural Similarity Index (SSIM)**: SSIM [140] is a widely used perceptual image quality metric that assesses the similarity between a reference image and a distorted image. The SSIM index is a decimal value ranging from 0 to 1, where 1 indicates a perfect similarity between the two images. The SSIM metric takes into account the luminance, contrast, and structural similarity between the images, making it more robust than traditional metrics like PSNR.

— **Multi-Scale Structural Similarity Index (MS-SSIM)**: MS-SSIM [139] is an extension of the SSIM metric that incorporates multiple scales to assess the structural similarity between images. MS-SSIM takes into account the perception of structural similarity at different levels, including global and local structural information. MS-SSIM provides a more comprehensive evaluation of structural similarity by considering information at multiple scales. It is commonly used in image and video quality assessment to capture perceptual differences that may not be captured by single-scale metrics like SSIM or PSNR.

— **Video Multimethod Assessment Fusion (VMAF)**: VMAF [80] is designed to assess the perceived quality of videos by considering various visual factors that influence human perception.

  VMAF takes into account a range of spatial and temporal features, including contrast, luminance, texture, and motion. It utilizes a machine learning algorithm that is trained on large-scale subjective quality datasets to predict human judgment of video quality. The output of VMAF is a score ranging from 0 to 100, where higher scores indicate better perceived quality.

— **FVVDP**: FovVideoVDP [96] is a video difference metric that simultaneously considers spatial, temporal, and peripheral aspects of perception. It addresses the complex interaction between spatial and temporal sensitivity in different retinal locations. Derived from psychophysical studies, the metric incorporates models for contrast

sensitivity, cortical magnification, and contrast masking.

Specifically, we investigate the consistency and uncertainty associated with these VQMs at a specific SUR threshold.

5.2.2 Experimental resuls

We compute the VQMs for a fixed SUR threshold, namely p%SUR$_{\text{emp}}$ (refer to Section 3.2.2 for the definision of p%SUR$_{\text{emp}}$) on VideoSet and AMZ-HDR-VJND dataset. We then analyze the consistency and uncertainty of different VQMs across these datasets for various video contents.

VideoSet 1080p

For VideoSet [138], the original VW-JND annotations for a given content J^m (refer to Eq.(3.1)) are provided in terms of QP values. We convert each element of J^m into its corresponding VQM scores and then calculate the p%SUR$_{\text{emp}}$ for each VQM. The PSNR in this study is computed on the Y channel. The VMAF version used in this work is v0.6.1 and for the FVVDP, we used v1.2.0. The parameters are set as $L_{\text{Peak}} = 165.8$, $contrast = 435$, $\gamma = 2.2$, $E_{\text{ambient}} = 100$, ppd[pix/deg] $= 60.8$, and $k_{\text{ref}} = 0.005$. The parameters are chosen following the recommenations in [91].

The results are presented in Table 5.1. We present the mean values of 80%SUR$_{\text{emp}}$, 75%SUR$_{\text{emp}}$, and 70%SUR$_{\text{emp}}$ across the 1080p of VideoSet in Table 5.1. The mean value of 75%SUR$_{\text{emp}}$ on VideoSet for VMAF is 93.62, which aligns with previous studies [115, 7] that suggest a first 75%SUR of approximately 94 for VMAF.

It can be observed that p%SUR$_{\text{emp}}$ for the first JND increases with higher values of p for all VQMs. Additionally, it is evident that for all these VQMs, the value of the VQM for p%SUR$_{\text{emp}}$ is highly content dependent.

Taking VMAF as example, for 75%SUR, the mean value is 93.62, with a minimum of 75.22 and a maximum of 99.97. This indicates that, for content A, if the compressed version has a VMAF score around 99.97, 25% (1-75%) of the population can already perceive the difference between this compressed version and the pristine version. However, for another content B, we need to degrade the quality until it reaches 75.22 so that 25% of the population can perceive the difference. This suggests that even though VMAF is highly correlated with human perception of quality, it is still not precise enough to measure the SUR in terms of JND.

Table 5.1 – Benchmark VQMs on 70%SUR, 75%SUR and 80%SUR on VideoSet 1080p for first JND

VQM (min-max)	Mean	Min	Max	COV	Avg 95%CI lower_b	Avg 95%CI upper_b	Avg 95%CI range	NAvg 95%CI range
80%SUR								
PSNR (23.4-60)	42.6826	32.4392	54.9418	0.0825	41.7589	44.7754	3.0165	0.0824
SSIM (0.7-1)	0.9947	0.9760	0.9998	0.0037	0.9931	0.9972	0.0041	0.0135
MS-SSIM(0.7-1)	0.9917	0.9662	0.9996	0.0053	0.9897	0.9950	0.0053	0.0186
VMAF (5.3-100)	94.5579	75.2246	99.9676	0.0399	92.6803	97.1454	4.4651	0.0471
FVVDP (4.8-10)	9.4835	8.5792	9.9705	0.0215	9.3727	9.6782	0.3056	0.0585
75%SUR								
PSNR (23.4-60)	42.2025	31.7613	53.3174	0.0823	41.2626	43.8400	2.5775	0.0704
SSIM (0.7-1)	0.9939	0.9689	0.9993	0.0043	0.9919	0.9964	0.0045	0.0149
MS-SSIM(0.7-1)	0.9907	0.9634	0.9990	0.0060	0.9881	0.9938	0.0057	0.0199
VMAF (5.3-100)	93.6156	75.2246	99.9676	0.0427	91.4264	96.2902	4.8638	0.0514
FVVDP (4.8-10)	9.4287	8.5529	9.9105	0.0231	9.3069	9.6044	0.2976	0.0569
70%SUR								
PSNR (23.4-60)	41.8065	31.1083	53.0536	0.0836	40.9990	43.3017	2.3027	0.0629
SSIM (0.7-1)	0.9932	0.968911	0.9990	0.0047	0.9912	0.9956	0.0045	0.0147
MS-SSIM(0.7-1)	0.9898	0.9634	0.9988	0.0065	0.9873	0.9929	0.0056	0.0194
VMAF (5.3-100)	92.7761	72.7199	99.9607	0.0473	90.6152	95.5342	4.9190	0.0519
FVVDP (4.8-10)	9.3795	8.4782	9.8811	0.0248	9.2667	9.5492	0.2825	0.0541

To measure the consistency of different VQMs in terms of SUR, we calculate the Coefficient Of Variation (COV) for p%SUR$_{\text{emp}}$. COV is the ratio of the standard deviation to the mean, serving as an indicator of variability. In this context, it is utilized because different VQMs operate on different scales. A larger COV indicates lower consistency for p%SUR$_{\text{emp}}$. Table 5.1 reveals that, for p values of 80, 75, and 70, SSIM exhibits the highest level of consistency among the six VQMs.

To visually represent the consistency of different VQMs, as shown in Figure 5.1, we plot the distributions using blue bars for the quality scores across the entire dataset (comprising 220 SRCs with QP values ranging from 1 to 51). Additionally, we use pink bars to represent the distributions of 75%SUR$_{\text{emp}}$ for each SRC. The y-axis uses a logarithmic scale.

The distribution of 75%SUR$_{\text{emp}}$ for PSNR appears relatively wide compared to the entire dataset. In contrast, the distributions for VMAF and FVVDP are relatively narrower. SSIM and MS-SSIM exhibit the narrowest distribution range for 75%SUR$_{\text{emp}}$, in line with its COV values presented in Table 5.1.

We also compute the 95%CI of p%SUR$_{\text{emp}}$ using the method introduced in Section 4.3.2, as presented in Table 5.1. We calculate the mean of the lower bound and upper bound for each VQM across the dataset. Notably, the lower bound and the upper

bound exhibit the same trend as the mean of $p\%\text{SUR}_{\text{emp}}$. To account for the varying scales, we normalized the CI range using the minimum and maximum values observed for each VQM across the entire VideoSet dataset:

$$Norm(95\%\text{ CI range}) = \frac{95\%CI_u - 95\%CI_l}{\max(VQM) - \min(VQM)}. \tag{5.1}$$

The mean of the normalized CI range is listed in Table 5.1 under the column 'NAvg 95%CI range'. Notably, SSIM exhibits the smallest CI range.

AMZ-HDR-VJND

Similar to VideoSet, we conducted a similar analysis on our collected HDR dataset: AMZ-HDR-VJND. There are two main differences compared with VideoSet. Firstly, the proxy of SUR for our collected datasets is Constant Rate Factor (CRF) instead of Quantization Parameter (QP) in VideoSet (see Chapter 3, Section 3.2.3). Secondly, because we used different display devices, the display and environment parameters for FVVDP are different from the ones used in VideoSet. Detailed parameters are set in Table 5.2. The minimum and maximum luminance of the display were measured using i1Profiler,

Table 5.2 – Display and environment parameters for FVVDP on our SONY displays in HDR mode

Parameters	Value
EOTF	PQ
Resolution	3840x2160
Diagonal size	55 inches
Max luminance	600 nits
Min luminance	0.16 nits
Ambient light	10 nits

while the ambient light was measured using a luxmeter. It's important to note that in Chapter 2, Section 2.2.1, we used different displays and environments. The parameters mentioned here are measured for the SONY display used in our AtHome subjective test, and the ambient light represents the average of different home environments.

However, it's important to acknowledge that we did not have strict control over the ambient light in the AtHome subjective test. Participants were instructed to turn off the lights and close the curtains, and the ambient light was measured at the beginning of the entire test session.

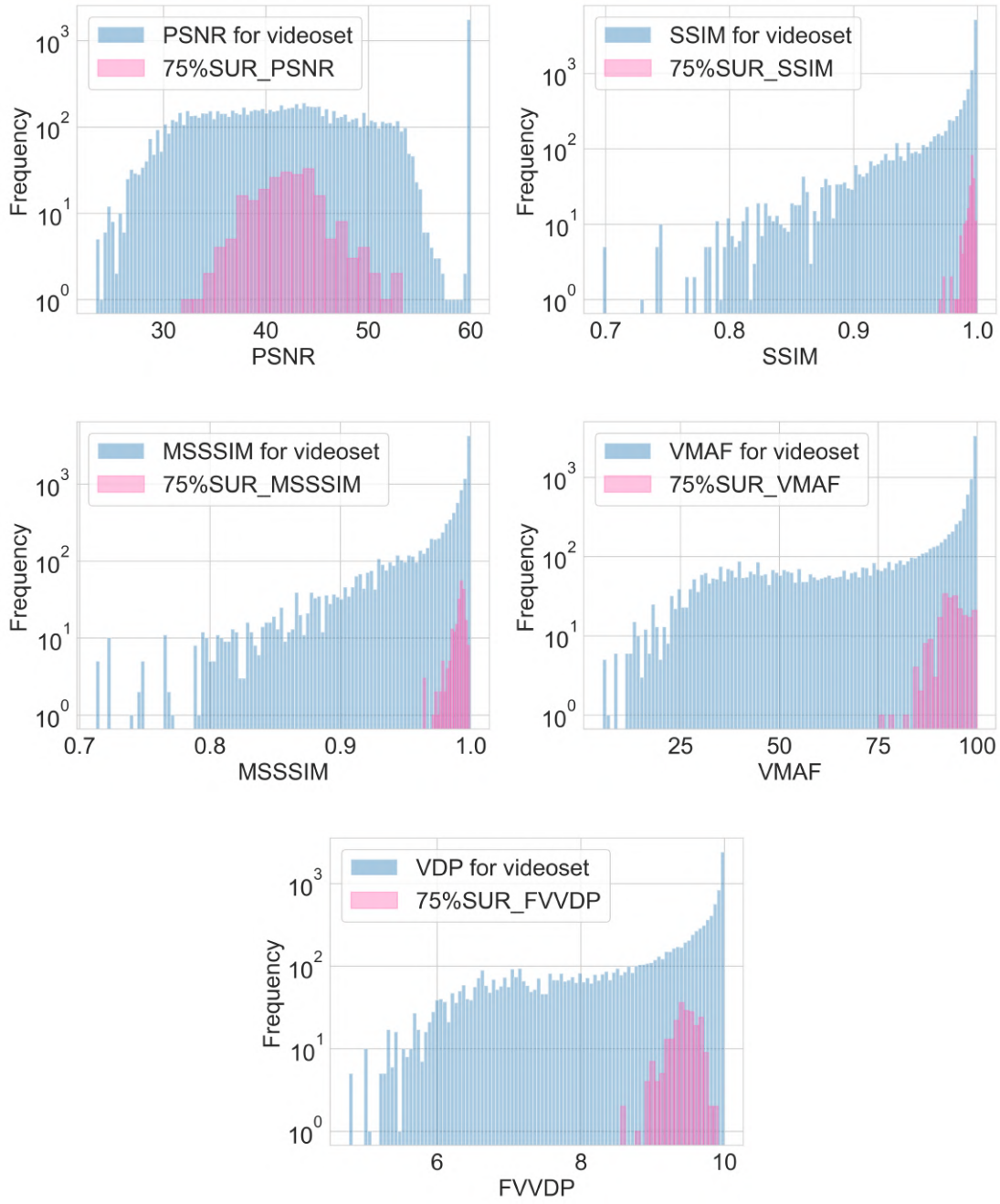


Figure 5.1 – Distributions of 75%SUR_{emp} and the distribution of VQMs on the entire datasets on VideoSet for different VQMs

The results are presented in Table 5.3. It can be observed that the mean value of 75%SUR_{emp} for PSNR, SSIM, MS-SSIM, VMAF, in AMZ-HDR-VJND is 49.33, 0.9987,

Table 5.3 – Benchmark VQMs on 75%SUR on AMZ-HDR-VJND for first JND

VQM (min-max)	Mean	Min	Max	COV	Avg 95%CI lower_b	Avg 95%CI upper_b	Avg 95%CI range	NAvg 95%CI range
75%SUR								
PSNR(37.2-72)	49.3318	38.0200	69.9868	0.0879	48.9674	50.4101	1.4427	0.0415
SSIM(0.98-1)	0.9987	0.9933	1.0	0.0010	0.9985	0.9991	0.0006	0.0517
MS-SSIM(0.95-1)	0.9948	0.9664	1.0	0.0050	0.9944	0.9958	0.0014	0.0304
VMAF(82.3-100)	96.9677	86.3406	100.0	0.0261	96.5740	97.8574	1.2835	0.0725
FVVDP(8.0-10.0)	9.4232	8.2501	9.9606	0.0225	9.3635	9.5619	0.1984	0.0977

0.9948 and 96.97, respectively. These values are all larger than those of VideoSet presented in Table 5.1. Only the mean value of FVVDP remains more or less the same in both datasets.

The Coefficients of Variation (COV) in AMZ-HDR-VJND are smaller than those of VideoSet, indicating that the VQMs are more consistent in our collected datasets. Additionally, the 95% CI range of the VQMs in AMZ-HDR-VJND is also smaller than those of VideoSet. It's important to note that it doesn't make sense to directly compare the Normalized Average 95% CI range of the VQMs across datasets because the JCPs are not designed in the same way.

Similar with VideoSet, we plot the histograms of the VQMs and $75\%\text{SUR}_{\text{emp}}$ in our collected dataset in Figure 5.2.

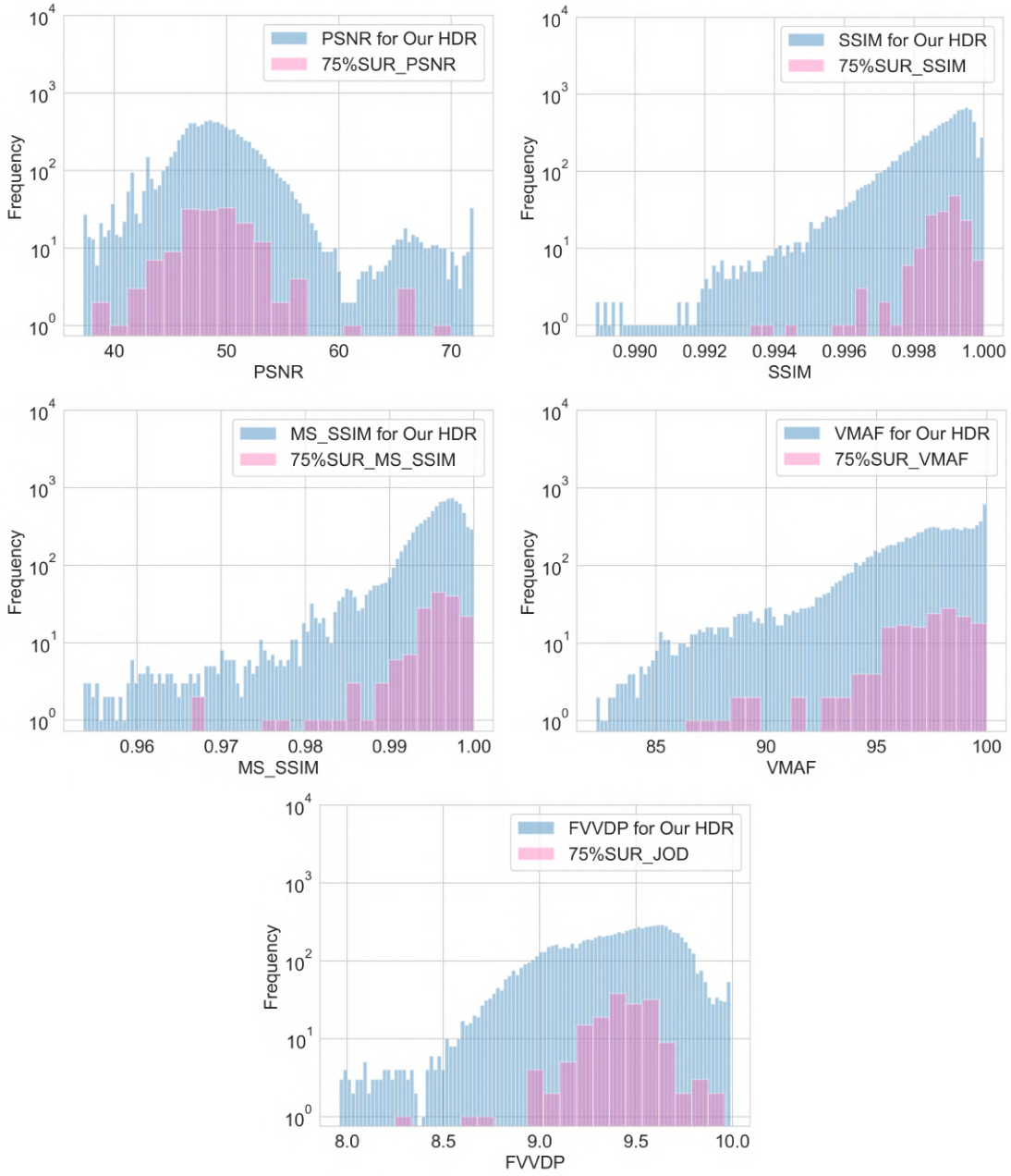
Experimental results on both datasets suggest that all these VQMs are highly content-dependent in terms of SUR. This indicates a significant challenge in developing VQMs that are not only highly correlated with human subjective opinion scores but also consistent and precise enough to measure SUR in terms of JND.

5.3 Prediction of SUR using VQMs as proxy

From the VQM resolving power analyses conducted in Section 5.2, it's evident that the VQMs towards SUR of JND are highly content dependent. As shown in Figure 5.3, using VMAF as an example, the $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ varies significantly across different video content in VideoSet. The $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ represents the transition from perceptually lossless to perceptually lossy coding of the VMAF proxy for $p\%$ of SUR value, as defined by Eq. (5.2) and Eq. (5.3).

$$\Delta\text{VMAF}_{\text{SUR}(p\%)} = \left|\text{VMAF}_{\text{SUR}(100\%)} - \text{VMAF}_{\text{SUR}(p\%)}\right|, \tag{5.2}$$

Figure 5.2 – Distributions of 75%SUR_{emp} and the distribution of VQMs on the entire datasets on AMZ-HDR-VJND for different VQMs



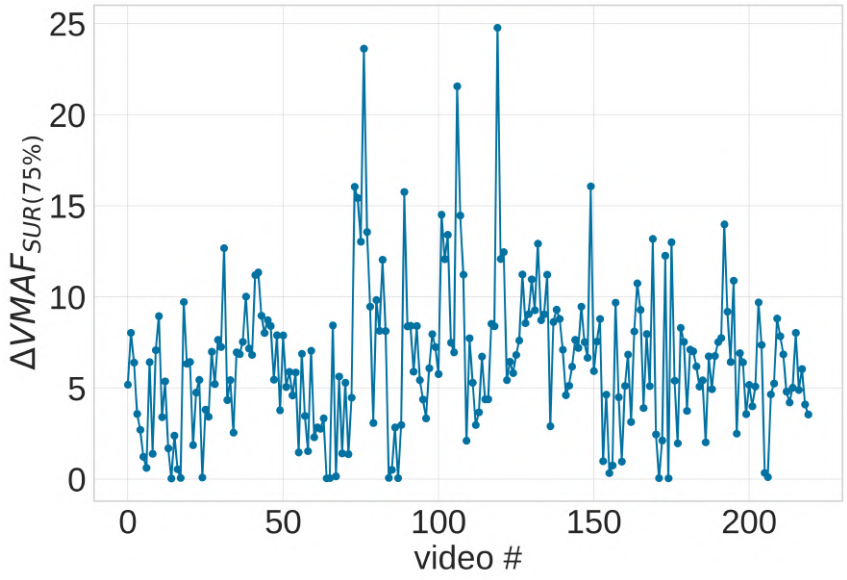


Figure 5.3 – $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ where the first JND occurs for for the $100 - p\% = 25\%$ ($p\% = 75\%$) of viewers in VideoSet [138] for 1080p.

where $\text{VMAF}_{\text{SUR}(p\%)}$ is the VMAF proxy where $p\%$ of viewers cannot see the transition and is defined as:

$$\text{VMAF}_{\text{SUR}(p\%)} = \arg\min_{x} \left|\text{SUR}(100 - x) - p\%\right|. \tag{5.3}$$

This observation leads to a research question: Can we predict the VQM towards SUR of JND at a given threshold, for example 75%SUR? The prediction of VQM towards SUR is crucial for constructing an optimized bitrate ladder. In Adaptive Bitrate Streaming (ABR) methods [17], video content is encoded at multiple bitrate-resolution pairs known as representations. These representations are used to construct a bitrate ladder [8], enabling the dynamic adjustment of video quality according to the viewer's available bandwidth and device type.

Traditionally, a fixed set of representations, such as the HLS bitrate ladder [12], is used for all video content. However, this "one-size-fits-all" approach may not be optimal for different types of videos. To address this, the per-title encoding approaches were introduced, where an optimized bitrate ladder is created for each video content, resulting in improved Quality of Experience (QoE). In per-title encoding [29, 8, 6], various encoding parameters, including resolution, frame rate, and others, are assessed by encoding the videos using all possible combinations of these parameters. Subsequently, an optimized

bitrate ladder is constructed by selecting representations from a convex-hull [148] based on the quality measurements of the encoded representations.

Selecting a subset of representations from the convex-hull is a crucial step in constructing an optimized bitrate ladder. This selection process involves considering various factors, such as available network bandwidth, device capabilities, and perceptual quality metrics like VMAF. While some methods focus on selecting representations based on the probability of clients requesting specific bitrate versions [123, 130], other approaches prioritize the selection of representations to minimize perceptual similarity [103] between the chosen representations. These methods aim to avoid including representations in the bitrate ladder that have similar perceptual qualities, as this redundancy may lead to inefficient resource utilization. If we can predict the VQM towards SUR of JND, we can then use the predicted VQM to construct the bitrate ladder to avoid this redundancy.

What do videos with the highest and lowest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ look like?

From Figure 5.3, we can see that the $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ varies from 24.88 to 0.01 for different contents. What do the videos with the highest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ and lowest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ look like? Please refer to Annex G.

In this section, we propose a method to predict the VMAF towards SUR of JND. VMAF is widely used in the industry due to its superior correlation with human perceptual quality [11]. It is frequently employed to evaluate the quality of video representations and guide the bitrate laddering process [67].

First, we introduce the existing method in the state of the art to predict the VMAF towards 75%SUR of JND. Then, we describe our proposed prediction pipeline incorporating different VQMs. Finally, we present the experimental results of the prediction of VQM towards SUR of JND and draw some conclusions.

5.3.1 State-of-the-art methods

In the existing literature, two sources provide recommendations for determining the VMAF towards SUR of JND in JND-based bitrate laddering. Jan Ozer [115], in a blog post, reports an interview with an unnamed Netflix employee who suggests a constant step size of 6 for $\Delta\text{VMAF}_{\text{SUR}(75\%)}$without empirical evidence to support this recommen-

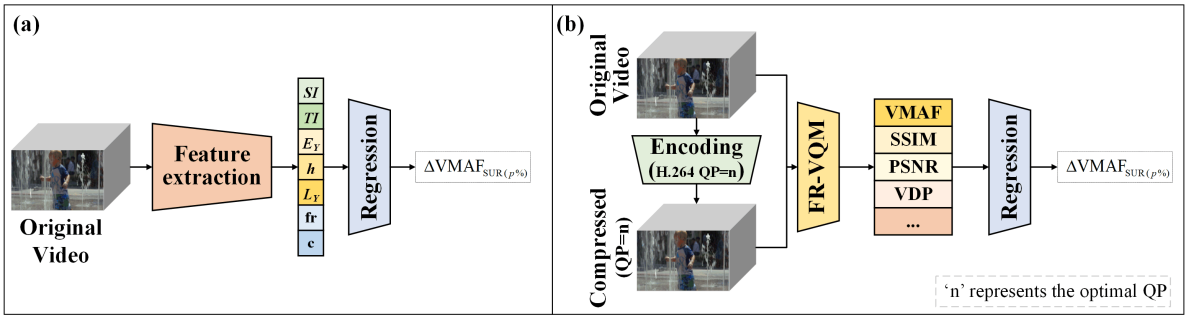


Figure 5.4 – Comparison of SOTA [7] (a), and our proposed pipelines (b) for $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ prediction. FR-VQM represents the full reference VQM.

dation. On the other hand, Kah *et al.* [66], in a publication, present the findings of a subjective study campaign on the acceptability and perceived quality of video content and propose a constant value of 2 for $\Delta\text{VMAF}_{\text{SUR}(75\%)}$. To further substantiate these two cases, Amirpour *et al.* [7] conducted an analysis by applying the two recommendations ($\Delta\text{VMAF}_{\text{SUR}(75\%)} = 2$ vs. 6) to VideoSet [138], comparing the outcomes with the subjective JND ground truth provided in the dataset.

In this study, each recommendation was interpreted as a simple linear model that predicts the $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ for each video content in the dataset. The results were subsequently analyzed using common error metrics. It was shown that using the constant value of the 6 rule for $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ performs significantly better than the constant value of 2.

However, these estimators still yield high error levels. The main reason for this discrepancy is the substantial variance of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ values across different video contents.

To effectively tackle the challenges in the current works, a content-specific framework [7] was developed to estimate $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ by utilizing features extracted from the reference video (see Figure5.4(a)). These features encompass the framewise features, including: (i) Spatial Information (SI) [58], (ii) Temporal Information (TI) [58], (iii) Spatial Energy (E) [104], (iv) Temporal Energy (h) [104], (v) Brightness (L), (vi) Colourfulness (c) [44], and (vii) Frame rate (fr).

By considering these features, the framework provided a more comprehensive estimation of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$, taking into account the spatial and temporal aspects, as well as brightness, colourfulness, and frame rate characteristics of the videos, as:

$$\Delta\text{VMAF}_{\text{SUR}(75\%)} = f(SI, TI, E, h, L, fr, c) \tag{5.4}$$

However, despite the utilization of this approach, the mean absolute error (MAE) was only reduced from 2.73 down to 2.11, in comparison to using the dataset's $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ mean of 6.93 as a constant estimator. Although an improvement was observed, the remaining error levels are still not satisfactory. Therefore, in this study, we investigate the use of VQMs to improve $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation accuracy.

5.3.2 $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ prediction pipeline using VQMs

Building upon the work of [7], we propose a new pipeline to further improve the estimation of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$. Illustrated in Figure5.4(b), we leverage VQMs on distorted video that is compressed with a fixed and optimal QP to enhance the estimation process. The rationale behind this idea is that quality metrics at a fixed QP provide information on how the VQMs interact with compression across different video content. By integrating VQMs into the $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation framework, we aim to achieve more refined and precise predictions of the perceptual differences in video content. The complete pipeline consists of two stages: *A)* computing VQMs and *B)* Regression on $\Delta\text{VMAF}_{\text{SUR}(75\%)}$based on VQMs on the optimal QP. Additionally, we conduct optimal QP selection to enhance the prediction accuracy and reduce the complexity of the pipeline.

We use the PSNR, SSIM, MS-SSIM, VMAF, FFVDP to enhance $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation. Detailed description of these VQM can be found in Section 5.2.1.

Optimal QP selection

Due to the computational cost associated with encoding videos at multiple QPs and measuring the corresponding quality metrics, we propose to encode the video at a single fixed QP (*qp*) and calculate the VQMs specifically at that optimal QP.

The ultimate optimal QP is determined using a regression model via a brute-force approach. Among the evaluated QPs, we identify the one yielding the lowest Mean Absolute Error (MAE) after regression. As depicted in Figure 5.5, illustrating the test set's MAE across various QPs utilizing a linear regression model with only VMAF as input, it becomes evident that the selection of QP notably influences the model's prediction accuracy. The results indicate that the lowest MAE is achieved at a QP value of 29 on VideoSet 1080p.

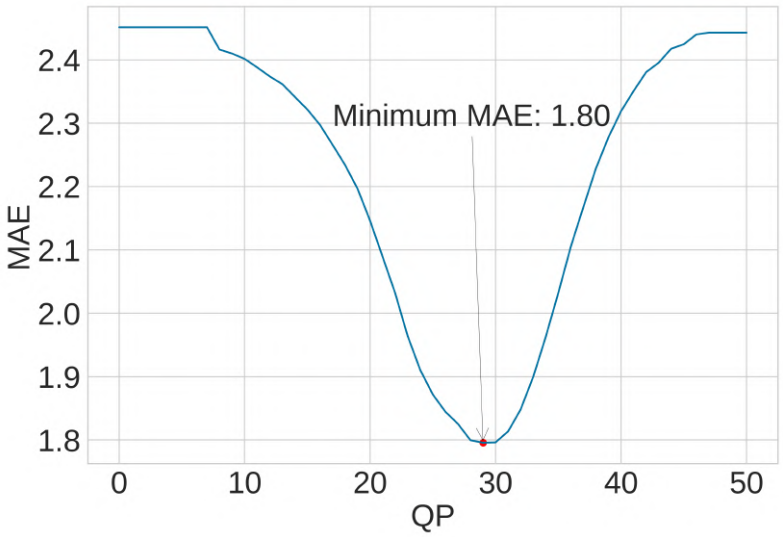


Figure 5.5 – MAE vs. QP for ridge regression with only VMAF as input features on VideoSet 1080p. We used an 80/20 train-test split over 20 random splits. The regularization parameter α is set to 0.5 for ridge regression.

Regression

Our prediction model for estimating $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ can be represented as:

$$\Delta VMAF_{SUR(75\%)} = f(\text{VMAF}_{qp}, \text{SSIM}_{qp}, \text{PSNR}_{qp}, \text{MS-SSIM}_{qp}, \text{VDP}_{qp}), \tag{5.5}$$

where qp is the optimal QP previously selected.

Initially, a foundational regression model, *i.e.,* ridge regression, with the complexity parameter $\alpha = 0.5$ was applied. Furthermore, a more optimized machine learning regression model, XGBoost (eXtreme Gradient Boosting) [24], was employed using parameters $n\_estimators = 100$, $max\_depth = 1$, and $booster = gbtree$.

To evaluate the contribution and dependency of these features in predicting the

$\Delta\text{VMAF}_{\text{SUR}(75\%)}$, we systematically eliminate the least important feature one by one until only one feature remains. We additionally conducted a comparative analysis of feature importance between the video complexity features outlined in [7] and the optimal VQM features for predicting $\Delta\text{VMAF}_{\text{SUR}(75\%)}$. Further details of the outcomes are provided in Section 5.3.3.

5.3.3 Experimental Results

The evaluation of the proposed method is conducted on VideoSet and AMZ-HDR-VJND. Prior to commencing the experiment, a preliminary data cleaning process was undertaken. As shown in Figure 5.6, the box plot of the $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ of raw data in VideoSet shows that there are few data points that deviate significantly from the majority of the dataset. These points have been removed to prevent these extreme values from distorting the overall analysis and interpretation of the data.

The remaining data points were divided into an 80% training set and a 20% test set. Train test split is conducted 20 times using 20 different random seeds. The reported results are the mean values obtained from these multiple splits.

The results are summarized in Table 5.4. When utilizing a fixed $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ of 2 [66], the MAE is 4.29. Increasing the fixed $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ to 6 [115] results in a reduced MAE of 2.59. The state-of-the-art method [7] achieves an MAE of 2.01. The results indicate that our proposed pipeline both ridge and XGBoost regression leads to lower MAE than SOTA. Utilizing XGBoost with VMAF_{28}, SSIM_{28} can further reduce the MAE to 1.67. The overall results presented in Table 5.4 demonstrate that incorporating quality metrics of video compressed with optimal QP can significantly reduce the MAE in $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ modeling.

Figure 5.7 portrays the feature importance of both the optimal VQM features and the video complexity features outlined in [7]. This importance is derived from the absolute values of the coefficients associated with each feature within the ridge regression model. The results clearly indicate that VMAF_{28} exerts the most substantial influence on the prediction model. Moreover, the VQM features demonstrate notably greater impact on $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ prediction compared to the video complexity features. This finding underscores the notion that VQMs not only inherently encompass video complexity features,

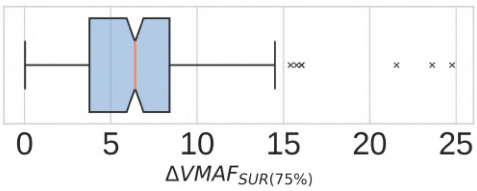


Figure 5.6 – The box plot of the $\Delta\text{VMAF}_{\text{SUR(75\%)}}$. Data points falling below the lower threshold ($Q_1 - 1.5IQR$) or exceeding the upper threshold ($Q_3 + 1.5IQR$) are considered outliers and have been excluded, as indicated by the 'x' marker.

Table 5.4 – Experiment results of different $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation models on VideoSet 1080p.

Model	Features	*qp*	MAE
Jan *et al.* [115]	$\Delta\text{VMAF}_{\text{SUR}(75\%)}= 2$	-	4.29
Kah *et al.* [66]	$\Delta\text{VMAF}_{\text{SUR}(75\%)}= 6$	-	2.59
Amirpour *et al.* [7]	SI, TI, E, h, L, c, fr	-	2.01
Ridge regression	vmaf, ssim, vdp, psnr, ms-ssim	30	1.77
Ridge regression	vmaf, ssim, vdp, psnr	30	1.77
Ridge regression	vmaf, ssim, vdp	29	1.78
Ridge regression	vmaf, vdp	29	1.78
Ridge regression	vmaf	29	1.80
XGBoost	vmaf, ssim, vdp, psnr, ms-ssim	28	1.73
XGBoost	vmaf, ssim, vdp, psnr	28	1.74
XGBoost	vmaf, ssim, vdp	28	1.73
XGBoost	vmaf, ssim	28	**1.67**
XGBoost	vmaf	28	1.72

but also furnish supplementary information, enhancing the predictive accuracy.

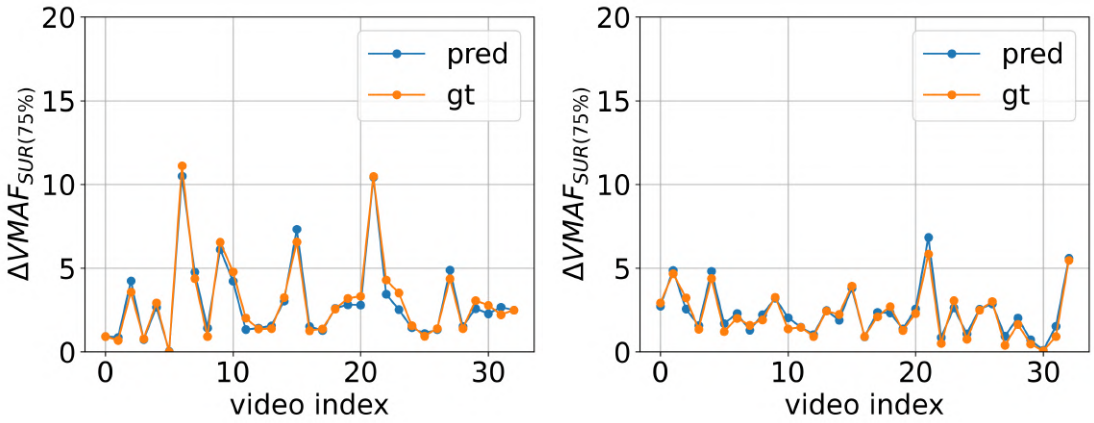
Figure 5.7 – Feature importance of optimal VQM features and video complexity features in [7]

Similarly, we employed the same pipeline on our collected HDR dataset. The results are summarized in Table 5.5. The optimal encoding parameter is CRF equals to 20. Surprisingly, the MAE in our collected HDR dataset is much smaller than in VideoSet. This might be due to the fact that our datasets used CRF as a proxy while VideoSet used QP. When encoding video in CRF mode, the encoder automatically adjusts the QP of each frame to maintain a constant quality level. Consequently, VQMs calculated at a fixed CRF value can help the model extrapolate the $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ more easily than those calculated at a fixed QP value. We visualized the prediction of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ using the Ridge regression model with VMAF, SSIM, PSNR as input. Similar to the approach

Table 5.5 – Experiment results of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation models on AMZ-HDR-VJND dataset.

Model	features	CRF	MAE
Ridge regression	vmaf	20	0.30
Ridge regression	vmaf, ssim, psnr	20	0.29
XGBoost	vmaf, ssim, psnr	20	0.40

used for VideoSet, we used 80% of the datasets as training set and the remaining 20% as test set, using different seeds for splitting the data. Figure 5.8 shows the predicted values and the ground truth for the test set with different seeds. The results for other seeds can be found in Annex H. It can be seen that the model effectively captures the trend of the ground truth across different splits.



(a) seed 0 (b) seed 1

Figure 5.8 – Visualization of the prediction of $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ using Ridge regression with VMAF, SSIM, PSNR as input features on AMZ-HDR-VJND dataset.

5.3.4 Discussion

In this section, we try to answer the research question: Can we predict the VQM towards SUR of JND at a given threshold, for example 75%SUR? Using VMAF as an example, we proposed a pipeline to predict the VMAF towards 75%SUR of JND. The pipeline consists of two stages: *A)* Computing VQMs and *C)* Regression based on VQMs on the optimal QP/CRF. We used the PSNR, SSIM, MS-SSIM, VMAF, FFVDP to enhance $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ estimation. The results show that our proposed pipeline can significantly reduce the MAE in $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ modeling. The optimal encoding parameter

is CRF 20 in our collected AMZ-HDR-VJND dataset and QP 28 in VideoSet 1080p. The MAE in our collected dataset is significantly smaller than in VideoSet, which might be because our collected datasets use CRF as a proxy instead of QP. The results indicate that incorporating quality metrics of compressed video with optimal QP/CRF can significantly reduce the MAE in $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ modeling.

However, in practice, accurately predicting the VQM value for SUR of JND at a given threshold, such as VMAF, does not directly inform us how to encode the video to achieve this VQM value. For instance, we need to encode the video at multiple QP/CRF settings to obtain the VMAF scores and then compare them to the target value to determine the appropriate QP/CRF to use. This process is computationally expensive. Therefore, in the next section, to improve efficiency and accuracy, we aim to directly predict the encoding parameters towards SUR of JND.

5.4 Prediction of SUR using encoding parameters as proxy

It is well known that subjective test is expensive and time-consuming, especially for VW-JND. Therefore, it is crucial to develop VW-JND prediction methods. In previous section 5.3, we proposed a pipeline to predict the SUR using VQM as proxy. In this section, we aim to predict the SUR using encoding parameter (*e.g.,* QP, CRF) as proxy, which is more efficient and practical for industry applications.

Wang *et al.* [137] proposed a model to predict SUR curve with QP proxy by using support vector regression (SVR) under the assumption that the individual JND points of a group users follow a normal distribution. Their model infers SUR values from VMAF [80] Quality Degradation Features concatenated with Masking Effect Features [50, 51] and is trained on VideoSet [138], the 75%SUR point can then be derived by the predicted SUR curve.

[135] is the extended work of [137], where the *2nd* and *3rd* JND points are predicted using 3 different settings in which the reference inputs of the predictor are different.

Instead of predicting SUR with encoding parameter QP as proxy, Zhang *et al.* [151] proposed a novel perceptual model to predict SUR versus bitrate, which is more widely used in practice. Three kinds of features, Masking features, re-compression features and basic attribute features, are extracted from original reference video to build a feature vector, which will be used to conduct a Gaussian Processes Regression (GPR) to predict

SUR.

Using deep learning, Zhang *et al.* [152] developed the Video Wise Spatial SUR method (VW-SSUR) for predicting the SUR with QP proxy for compressed video along with the Video Wise Spatial-Temporal SUR (VW-STSUR) to boost the prediction accuracy.

Nami *et al.* [109] proposed a multi-task deep learning framework to predict both PW-JND and VW-JND. The framework jointly learns the three JND levels (first, second, and third JND), the visual attention map with one JND level, and the visual attention map with all three JND levels. They don't predict the SUR curve but the QP value of 75%SUR on the SUR curve.

However, previous works assume that the individual VW-JND of a group viewers follows Gaussian distribution, which might not be the optimal modeling method. Besides, when predicting SUR curve, previous works are computationally expensive because they extract features from SRC and from every encoded PVS and predict the individual SUR score of each PVS to derive the SUR curve. Therefore, we first investigate the modeling of SUR curve and then propose a novel SUR prediction model only based on SRC.

5.4.1 Parameter-driven model

We firstly investigated the modeling of the group-based VW-JND rather than using a simple normality test as in the previous works [138, 136], in order to find the mathematical model that best fits SUR. Afterwards, we proposed a SUR prediction method via predicting the model parameters obtained by the modeling fit, which is called parameter-driven model, in preference to the commonly used point-by-point models. The entire pipeline is shown in Figure 5.9.

Modeling of SUR

When modeling the SUR of VW-JND, previous works [138, 136] conducted Jarque-Beta test [62] to verify the normality of the VW-JND position of every subject. However, when revisiting the original distribution of VW-JND annotations, we found that Gaussian distribution is not necessarily the best modeling of the VW-JND distribution as shown in Figure 3.2 in Chapter 3. Therefore, based on the definition of the empirical SUR curve and p%SUR in Section 3.2.2, we set the Complementary Cumulative Distribution Function (CCDF) of different candidate distributions (*e.g.,* Gaussian, Sigmoid, Weilbull, *etc.*) as model functions to find the best fit function of the empirical SUR curve. The model

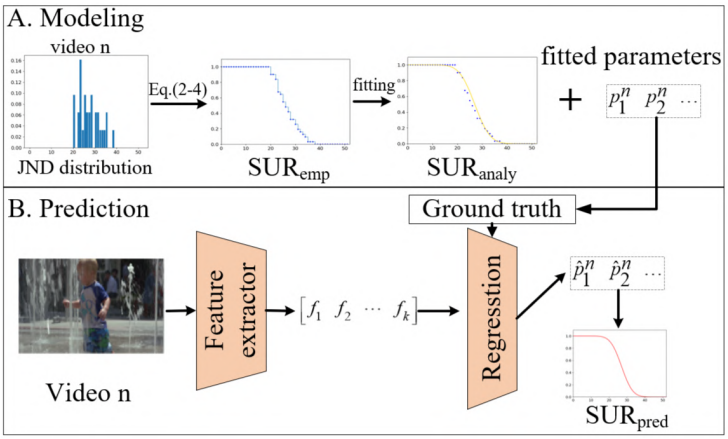


Figure 5.9 – Illustration of the pipeline of SUR and JND modeling (A) and prediction (B)

Table 5.6 – Summary of candidate model functions. (NB para is the number of parameters in model function)

Name	Model function	NB para
Polynomial-3	$f(x) = \sum_{k=0}^{n} a_k x^k$	4
Polynomial-4		5
Gaussian	$1 - \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$	2
2-para-logistic	$1 - \frac{1}{1+e^{-(x-\mu)/s}}$	2
4-para-logistic	$f(x) = b + \frac{L}{1+e^{-k(x-x_0)}}$	4
Weibull	$e^{-\left(\frac{x}{\lambda}\right)^k}$	2
Gumbel	$1 - e^{-e^{-(x-\mu)/\beta}}$	2
Rayleigh	$e^{\frac{-x^2}{(2\sigma^2)}}$	1

functions were named as analytical SUR. We also proposed a criteria to select the best fit function of the empirical SUR curve.

After computing the empirical SUR from the VW-JND annotations, we fitted the discrete points in empirical SUR with 8 model functions (Table 5.6) for each video. It can be easily proved that the CDF (Eq.(3.3)) of a distribution is monotonic non-decreasing, thus SUR (Eq.(3.4) case 1 using QP proxy) is monotonic non-increasing. Therefore, monotonic constraint was applied during least-squares optimization for polynomial model function.

In addition to MAE and RMSE, we use $\Delta p\%\text{SUR}_{|\text{E}-\text{A}|}$ (Eq.(5.6)) to evaluate the

candidate model functions, where $p\%$ is set to 75%.

$$\Delta p\%\text{SUR}_{|\text{E}-\text{A}|} = |p\%\text{SUR}_{\text{emp}} - p\%\text{SUR}_{\text{analy}}| \tag{5.6}$$

VideoSet [138] was used to evaluate the modeling of SUR and JND. We used the individual VW-JND annotations of each SRC to generate the SUR_{emp} and $75\%\text{SUR}_{\text{emp}}$ (Eq.(3.4)-(3.5)) and every discrete points of SUR_{emp} were used to fit the model functions listed in Table 5.6. Scikit-learn linear regression was used for polynomial fittings and monotonic constraints were applied by using Polyfit [1]. Non-linear least squares from SciPy were used for other model functions.

Table 5.7 – Mean of MAE, RMSE and $\Delta 75\%\text{SUR}_{|\text{E}-\text{A}|}$ for different model functions with VideoSet [138]

Name	MAE	RMSE	$\Delta 75\%\text{SUR}_{\|\text{E}-\text{A}\|}$
Polynomial-3	0.1204	0.1466	5.0614
Polynomial-4	0.1085	0.1338	4.7420
Gaussian	0.0147	0.0253	0.6625
2-para-logistic	0.0156	0.0250	0.5875
4-para-logistic	0.0164	**0.0236**	**0.5761**
Weibull	**0.0138**	0.0240	0.6761
Gumbel	0.0220	0.0343	0.5977
Rayleigh	0.1451	0.1703	8.9114

For every SRC in the 4 resolutions of VideoSet ($220 \times 4 = 880$ SRC in total), we calculated the MAE and RMSE between SUR_{emp} and $\text{SUR}_{\text{analy}}$ and also the difference between empirical and analytical 75%SUR : $\Delta 75\%\text{SUR}_{|\text{E}-\text{A}|}$ (Eq.(5.6)) with different fitting functions. The results are shown in Table 5.7. It can be observed that the CCDF of Gaussian distribution is not the best modeling for SUR. 4-para-logistic model function outperforms the other candidate model functions both in RMSE and $\Delta 75\%\text{SUR}_{|\text{E}-\text{A}|}$. This result is consistant with the results of MLE in Section 4.3.3 in Chapter 4.

Prediction of SUR

We revisited the SUR prediction model proposed by Wang *et al.* [137] namely the baseline. Furthermore, we analysed the two main drawbacks of the baseline model and proposed solutions to solve these issues.

1. https://github.com/dschmitz89/Polyfit

As shown in Figure (5.10), the input of the baseline model is the uncompressed source video (SRC). SRC is firstly compressed with different encoding parameters (*e.g.*,QP 1 to 51) to get a series of PVS (Processed Video Sequence). Afterwards, SRC and all PVS are segmented to small video patches both spatially and temporally to extract features from the eye fixation level. Two types of features are extracted from the segmented patches: Masking effect and Quality degradation (M and Q in Figure (5.10)). Masking effect is a measure of the spatial and temporal randomness [50, 51]. A high level of randomness masks distortions from the human eye, making it difficult to perceive the difference between the SRC and the distorted PVS. Quality degradation is calculated based on the difference of quality scores (*e.g.*, VMAF) between SRC and PVS. The masking effect and quality degradation feature vectors of one video are the histogram and cumulative curve of its video patches, respectively. When extracting the Quality degradation features, only video patches with significant quality degradation were selected to compute the final feature vector. The two feature vectors for SRC and each PVS are concatenated and are used to predict SUR scores by regression.

The baseline model is computational expensive as individual SUR scores of each PVS of a SRC must be computed to derive the SUR curve prediction. Accordingly, features from a SRC and its PVSs (*i.e.,* 52 sequences) have to be computed. This is the first drawback of the baseline model. The second drawback is that the SUR_{pred} curve of baseline model is not monotonic non-increasing because every individual point of SUR curve is predicted separately. The basic meaning of SUR is: for a given distortion level x, its SUR value is the percentage of subjects that are satisfied, *i.e.,* the percentage of subjects that do not perceive the difference between the reference video and all the distorted video whose distortion level is less than x. Therefore, it is not reasonable that when the distortion level increases, the SUR value increases.

To address these issues, we proposed a straightforward solution with the help of the modeling parameters of SUR. The pipeline of the proposed method is shown in Figure 5.9. Instead of predicting SUR scores of every PVS and getting the SUR curve accordingly, the parameters which describe the SUR curve (e.g., σ and μ for Gaussian) are predicted. Only SRC is used for prediction, hence these models are called **SRC-based** model. Masking effect features are extracted from SRC and Support Vector Regression (SVR) [128] is used for regression. Although SRC-based models are preferred in real-life applications, it is still interesting to understand how important the quality degradation information from PVSs is to the prediction of SUR and JND. Therefore, we also investigated **SRC+PVS-**

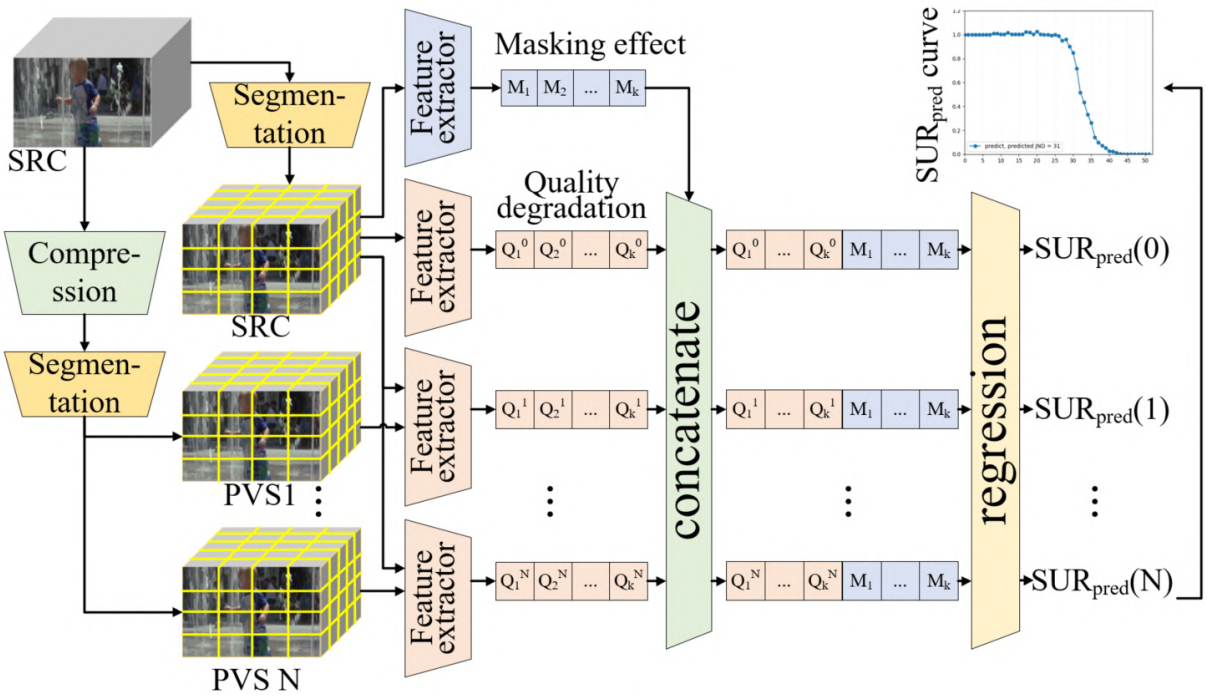


Figure 5.10 – Illustration of baseline SUR prediction model

based model, where masking effect and quality degradation features of every PVS are concatenated into one vector for regression to predict the modeling parameters.

Each prediction model was evaluated on videos using 5-fold cross validation. Radial basis function kernel was used for SVR. We firstly compare the baseline model (in-house implementation) and 3 SRC-based parameter-driven models with different model functions. Figure 5.11 shows the SUR prediction results of 2 SRCs. The dashed lines in orange, green and blue are the analytical SUR curves obtained from fitting to the red points in empirical SUR with Gaussian, 2-p-logistic and 4-p-logistic respectively. Plain lines with dots are the predicted SUR of the 3 SRC-based models obtained by predicting the parameters of the corresponding dashed lines. The green line is the SUR prediction of baseline model. Difference between **P**redicted and **A**nalytical SUR (denoted $\Delta SUR_{|P-A|}$) is the MAE between them for the QP/CRF values present in the datasets. $\Delta SUR_{|P-A|}$ indicates the error between ground truth (analytical SUR curve) and prediction SUR curve, but the modeling error (between empirical and analytical SUR curve) are not considered. Therefore, difference between **P**redicted and **E**mpirical SUR (denoted $\Delta SUR_{|P-E|}$) is evaluated as well. Δ75%SUR is evaluated in the same way. The results are shown in Table 5.8. The 3 SRC-based parameter-driven models outperform the baseline model both in ΔSUR

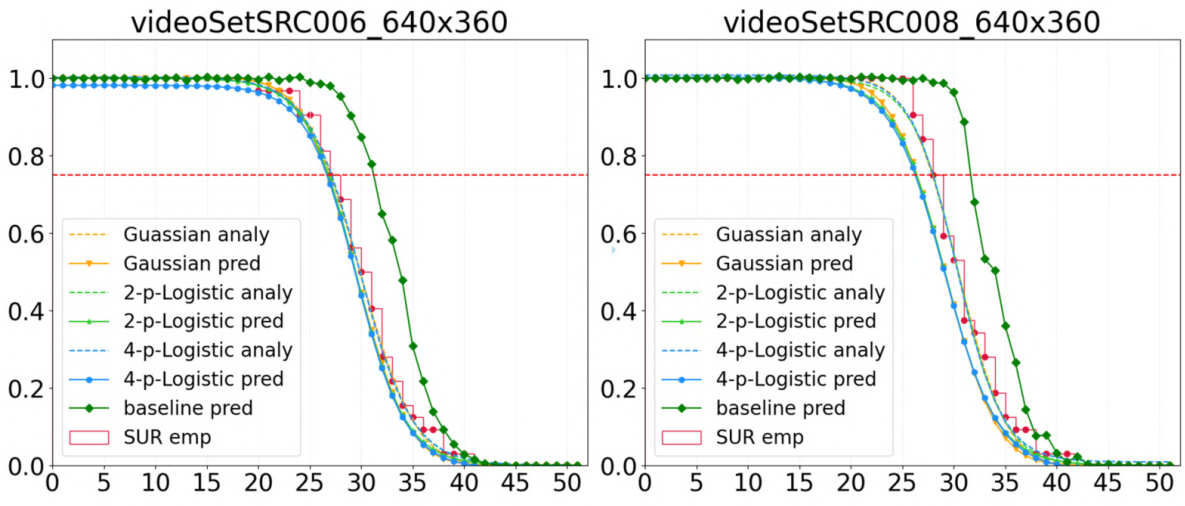


Figure 5.11 – Examples of SUR prediction results comparison between SRC-based models and baseline model

and Δ75%SUR. The prediction errors between Gaussian and logistic parameter-driven models are quite close. However, the 4-para-logistic which has the smallest modeling error performs worse than Gaussian and 2-para-logistic.

Table 5.8 – Averaged prediction error comparison between baseline model and 3 SRC-based parameter-driven models on VideoSet.

RES	Model name	ΔSUR		Δ75%SUR	
		$\|P-A\|$	$\|P-E\|$	$\|P-A\|$	$\|P-E\|$
360p	baseline	0.0769	0.0799	4.3682	4.3773
	2-p-Gaussian	**0.0459**	**0.0480**	2.4773	2.5864
	2-p-Logistic	0.0462	0.0489	**2.4455**	**2.5682**
	4-p-Logistic	0.0496	0.0515	2.4591	2.5909
540p	baseline	0.0786	0.0812	4.3182	4.2909
	2-p-Gaussian	**0.0397**	**0.0428**	2.1182	2.1045
	2-p-Logistic	0.0398	0.0437	**1.9727**	**2.0955**
	4-p-Logistic	0.0435	0.0458	2.0045	2.1000
720p	baseline	0.0783	0.0820	4.2864	4.2909
	2-p-Gaussian	**0.0433**	**0.0447**	**2.1636**	**2.2045**
	2-p-Logistic	0.0435	0.0459	**2.1636**	2.2364
	4-p-Logistic	0.0467	0.0476	**2.1636**	2.2318
1080p	baseline	0.0801	0.0834	4.6000	4.5591
	2-p-Gaussian	0.0412	**0.0431**	2.3455	2.2136
	2-p-Logistic	**0.0409**	0.0440	**2.1182**	**2.1773**
	4-p-Logistic	0.0439	0.0455	2.1455	2.1727

We also compared SRC-based model and SRC+PVS-based model, the results are

shown in Table 5.9. It can be observed that adding quality degradation information from PVSs will improve the prediction of both SUR and 75%SUR, but with a cost of encoding SRC to 51 PVSs.

Table 5.9 – Averaged prediction error comparison between SRC-based and SRC+PVS based model on VideoSet 1080p with Guassian modeling.

Model	ΔSUR		Δ75%SUR	
	$\|P-A\|$	$\|P-E\|$	$\|P-A\|$	$\|P-E\|$
SRC-based	0.0412	0.0431	2.3455	2.2136
SRC+PVS-based	**0.0377**	**0.0412**	**2.0727**	**2.1409**

Summary

In section, we proposed a novel pipeline for SUR modeling and prediction to predict the optimal encoding parameter only from SRC. Experiment results show that the proposed parameter-driven model (2-p-Logistic for instance) improves the mean SUR prediction error to 0.046, reducing it by 43.64% compared with the baseline and reduces the mean 75%SUR prediction error from 4.38 QP (baseline) to 2.27 QP. Furthermore, compared with SRC-based model, the SRC+PVS-based model slightly improves the mean prediction error of SUR curve and 75%SUR by 0.0019 and 0.0727 QP respectively, which means the quality degradation features from PVSs are not crucial to SUR prediction.

5.4.2 Further improvement of the prediction

In the previous Section 5.4.1, we firstly compare several different mathematical modelings of SUR curve and secondly compute the SUR curve by predicting the modeling parameters only based on the features of SRC. Nevertheless, the prediction errors of SUR curve and 75%SUR are still non-negligible due to the limitation of masking effect features. In this section, we propose a new SUR prediction framework based on the previous one that extract many different types of features other than the masking effect features followed by features selection and regression.

Similar with the SUR prediction framework in Figure 5.9 in Section 5.4.1, there are two steps (see Figure 5.12) for the entire pipeline: (1) Modeling; (2) Prediction. Modeling includes computing the empirical SUR curve (SUR_{emp}) from the JND distribution of the group-users and finding the best mathematics model to fit the SUR_{emp}. The fitted SUR curve is denoted as SUR_{analy}. After generating ground truth from modeling, we use SRC as

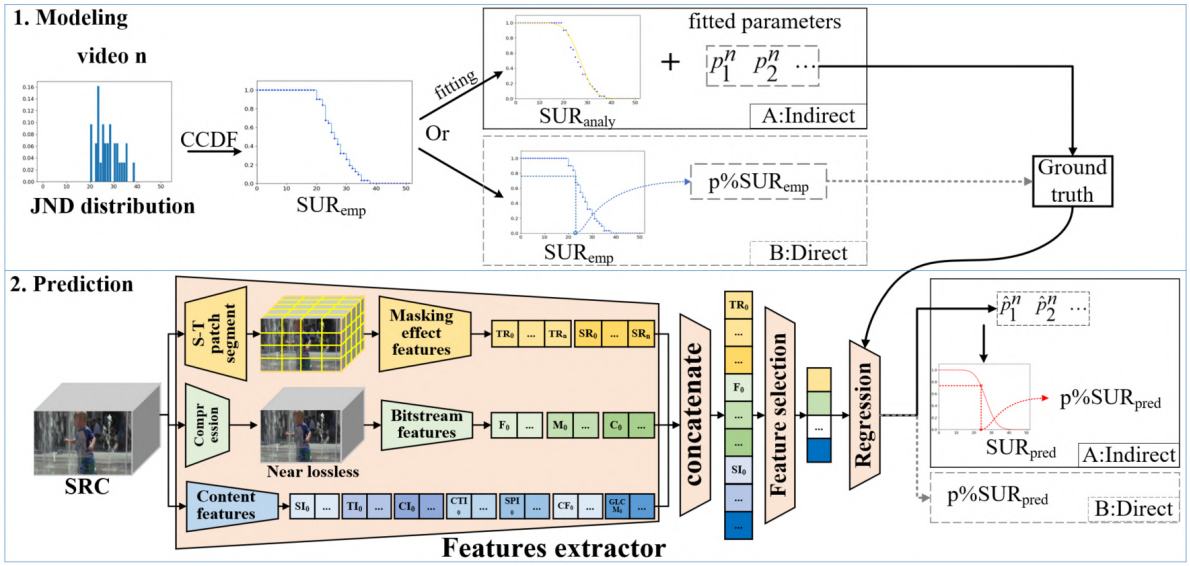


Figure 5.12 – Illustration of the pipeline of SUR and JND (1) modeling and (2) prediction framework (improved version of Figure 5.9)

input to extract features, select features and make predictions. The prediction framework is detailed as follows:

Feature extraction

Three types of features: (1) masking effect features, (2) bitstream features and (3) content features are extracted as illustrated in Figure 5.12.

Masking effect Features measures randomness/regularity temporally (temporal randomness (TR) [51]) and spatially (spatial randomness (SR) [50]). When the randomness is high, it will be difficult for human to perceive difference, as it masks the distortion for the HVS. Masking effect features were used in [137, 159] to predict JND.

As shown in Figure 5.12, SRC is segmented into small video patches both spatially and temporally to extract features from the eye fixation level. The dimensions of video patches are set the same as [137]. SR and TR are calculated on each small video patch to obtain feature matrices F_{SR} and F_{TR} (Please refer to Annex I for details). The statistic histogram (Eq. (5.7)) with number of bins equals to 20 is applied as pooling method to reduce the feature dimension.

$$\overrightarrow{SR} = Hist_{20}(F_{SR}), \quad \overrightarrow{TR} = Hist_{20}(F_{TR}) \tag{5.7}$$

Bitstream Features are widely used for light-weight quality estimation [121]. Before extracting bitstream features in Table 5.10, SRCs are first compressed into a near lossless PVS with CRF = 5 for our datasets. The bitstream features are extracted using videoparse [122] without decoding pixel information.

The temporal and spatial pooling function are defined as:

$$F_{time} = \{Mean, Std, Max, Skew, Kurt\} \tag{5.8}$$

$$F_{space} = \{Mean, Std\} \tag{5.9}$$

where $Mean$ is the average value, Std indicates standard deviation, Max denotes maximum, $Skew$ represents skewness, and $Kurt$ is the kurtosis. The dimension of features equals to the product of the dimension of F_{time} and F_{space} (*e.g.,* for motion features, we first compute the $Mean$ and Std of motions intra frame (spatially); afterwards, $Mean$, Std, Max, $Skew$ and $Kurt$ are calculated based on the two previous-computed spatial value ($Mean$ and Std for each frame) inter frame (temporally) respectively.)

Table 5.10 – Bitstream Features Summary

Features	dimension
Average framerate	1
Bitrate	1
Ratio(non − I) $= \frac{Nb(\text{non−I frame})}{Nb(\text{all frame})}$	1
Max(Framerate)	1
F_{time}(non - I frame size)	5
F_{time} {F_{space}(horizontal motion)}	5*2 = 10
F_{time} {F_{space}(vertical motion)}	5*2 = 10
F_{time} {F_{space}(motion)}	5*2 = 10
F_{time}(Temporal Complexity [122] per frame)	5
F_{time}($Spatial$ Complexity [122] per frame)	5

[1] Nb: number;
[2] F_{time} and F_{space} are the temporal (Eq. (5.8)) and spatial (Eq. (5.9)) pooling function.

Content Features include 7 types of features: Spatial Information(SI) [61], Temporal Information (TI) [61], Chrominance Information (CI) [146], Contrast Information (CTI) [146], Spatial Perceptual Information (SPI) [146], Colorfulness (CF) [44] and Grey Level Co-occurrence Matrix (GLCM) [42]. As illustrated in Figure 5.12, they are extracted directly from the pixel level of the SRC [88]. The temporal pooling function for content

features is the same with bitstream features (Eq. (5.8)), and the spatial pooling function is defind as:

$$F_{space} = \{Mean, Std, Max, Skew, Kurt\}. \tag{5.10}$$

The co-occurrence matrix (CM) is computed based on image patches. For each small patche, we calculated 6 features as shown in Eq. (5.11):

$$\begin{aligned} F_{patch} = \{&contrast, dissimilarity, homogeneity, \\ &ASM, energy, correlation\}, \end{aligned} \tag{5.11}$$

where $contrast = \sum_{i,j=0}^{l-1} CM_{i,j}(i-j)^2$, the dissimilarity $diss = \sum_{i,j=0}^{l-1} CM_{i,j}\,|i-j|$, the homogeneigy $homo = \sum_{i,j=0}^{l-1} \frac{CM_{i,j}}{1+(i-j)^2}$, Angular Second Moment: $ASM = \sum_{i,j=0}^{l-1} CM_{i,j}^2$, $energy = \sqrt{ASM}$ and $correlation = \sum_{i,j=0}^{l-1} CM_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2\sigma_j^2}}\right]$, in which l is the level of luminance of original image patch (l=255 for 8 bit image), i, j are the horizontal and vertical index of CM respectively; μ, σ are the mean and variance of CM.

Features selection

As shown in Figure 5.12, all the extracted features are concatenated into one vector. The exhibited vector has dimension of 399. We then used Forward-Sequential Feature Selection (F-SFS) [37] to select the optimal set of feature for SUR prediction. It is a greedy procedure. More specifically, we initially find the feature that minimize the cross-validated prediction error for the modeling parameters in indirect mode A or the p%SUR$_{\text{emp}}$ in direct mobe B, when an estimator is trained on this single feature. Once that first feature is selected, we repeat the procedure by adding the new feature that maximizes the cross-validated score to the set of selected features. The procedure stops when the desired number N of selected features is reached. Grid search was adapted to determine N.

Regression

The selected features will be fed into a SVR for prediction. As shown in Figure 5.12, there are two ways to predict p%SUR:

— A: indirect p%SUR prediction through SUR modeling

— B: direct p%SUR prediction without modeling

For the indirect mode, analytical SUR curve (SUR$_{\text{analy}}$) and its parameters are determined by fitting the empirical SUR curve (SUR$_{\text{emp}}$). The fitted parameters serve as

Table 5.11 – Content Features Summary.

Features	dimension
$SI^+ = F_{time}\{F_{space}\{\text{Sobel}[\text{Y}_n(i,j)]\}\}$	5*5 =25
$TI^+ = F_{time}\{F_{space}[M_n(i,j)]\}$ where $M_n(i,j) = Y_n(i,j) - Y_{n-1}(i,j)$	5*5 =25
$CI_{Cb} = F_{time}\{F_{space}[Cb_n(i,j)]\}$ $CI_{Cr} = F_{time}\{W_R \times F_{space}[Cr_n(i,j)]\}$ where $W_R = 1.5$	5*5+5*5=50
$CTI = F_{time}\{F_{space}[Y_n(i,j)]\}$	5*5 = 25
$SPI_{SI13} = F_{time}\{F_{space}[R_n(i,j)]\}$ where $R_n(i,j) = \sqrt{H_n(i,j)^2 + V_n(i,j)^2}$, $SPI_{HV13} = F_{time}\left\{\frac{mean[HV(i,j)]}{mean[\overline{HV}(i,j)]}\right\}$, $rg = R_n(i,j) - G_n(i,j)$, $yb = \frac{1}{2}(R_n(i,j) + G_n(i,j)) - B_n(i,j)$	5*5+5 = 30
$CF = F_{time}\{CF_n\}$ where $CF_n = \sigma_{rgyb} + 0.3\mu_{rgyb}$ $\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}$, $\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$	5
$GLCM = F_{time}\{F_{space}[F_{patch}(CM]\}$ where CM is the co-occurrence matrix[1]	5*(5*6)=150

[1] https://scikit-image.org/docs/0.7.0/api/skimage.feature.texture.html

[2] F_{time} and F_{space} are the temporal (Eq. (5.8)) and spatial (Eq. (5.10)) pooling function; F_{patch} is the functions to compute the texture features of the spatial patch with size $n = 64 * 64$.

[3] Y, Cr and Cb are the luminance and two chroma components; R, G and B are the red, green and blue channels.

ground truth for the regressor. We first obtain the SUR_{pred} curve by predicting the fitted parameters from the features. p%SUR can be computed from the SUR_{pred} curve.

For the direct mode, the ground truth for the regressor is $p\%\text{SUR}_{\text{emp}}$. This mode is useful if one is only interested in the specific value of p%SUR (*e.g.,* the demand of a streaming service provider is to satisfy 75% clients), but not the SUR curve, the direct prediction without modeling can be adapted as illustrated in the dotted box in Figure 5.12.

Experiments and results

The logistic function showed best prediction performance in the previous Section 5.4.1, hence it is employed in the indirect prediction model. Before feeding the extracted features to the SVR, all the features are normalized by applying z-score transformation. The estimator for F-SFS is the SVR and the metric of features selection is the Mean Square

Error (MSE) of the fitted paramters for the indirect mode and the $p\%SUR_{emp}$ for the direct mode. The optimal number of selected features is 55, detailed in Table 5.12. It can be observed that bitstream features has highest selection rate, which means the bitstream features such as motions are significant for SUR value prediction.

Table 5.12 – Number of features selected per category

Features type	Selected/original	Ratio of selection
Masking effect	9/40	0.2250
Bitstream	15/49	**0.3061**
Content features	31/310	0.1000

Each model is evaluated in our AMZ-HD-VJND datasets with 5-fold cross validation with random split (fixed random state). Hyper parameters of SVR are determined by grid search (kernel='rbf', C=0.1, epsilon=0.0001, gamma='scale'). Difference between **P**redicted and **A**nalytical SUR curve ($\Delta SUR_{|P-A|}$) is the Mean Average Error (MAE) between them. $\Delta SUR_{|P-A|}$ indicates the error between the fitted analytical SUR curve and the predicted one. Difference between **P**redicted and **E**mpirical SUR curve ($\Delta SUR_{|P-E|}$) is evaluated as well. $\Delta 75\%SUR$ is evaluated in the same way. The results are shown in Table 5.13. For the model who predict directly the $p\%SUR$ (Direct mode), the $\Delta SUR_{|P-A|}$ and $\Delta SUR_{|P-E|}$ don't exist. Similarly, we cannot compute $\Delta 75\%SUR_{|P-A|}$ because the SUR_{analy} doesn't exist without modeling.

Experiment results show that our proposed models, both direct and indirect mode, outperforms the basic parameter-driven model presented in Section 5.4.1, reducing $\Delta SUR_{|P-E|}$ by 40%. The direct $p\%SUR$ prediction mode without modeling has the smallest prediction error in terms of $\Delta 75\%SUR$. However, the indirect model provides us more information (the SUR curve and the 75%SUR value) compared to the direct model that outputs only the 75%SUR value. Furthermore, it could be observed in Table 5.13 that the indirect model has smaller standard deviation than direct $p\%SUR$ prediction model which indicates that the indirect model helps stabilize the variation of the prediction error.

Similarly, we applied the proposed framework on AMZ-HDR-VJND datasets. In addition to the bitstream features, we included SI TI 10bits [2] and WCG [3] features [73] for HDR content. The results, shown in Table 5.14, indicate that, similar to the AMZ-HD-VJND datasets, the direct $p\%SUR$ prediction mode without modeling has the smallest prediction

2. https://github.com/VQEG/siti-tools, last access: May 30, 2024.
3. https://github.com/junghyuk-lee/WCG-content-characterization, last access: May 30, 2024.

Table 5.13 – Average and variance of prediction error in AMZ-HD-VJND datasets

Model		ΔSUR		Δ75%SUR	
		|P − A|	|P − E|	|P − A|	|P − E|
Basic model (Section 5.4.1)	mean	0.1121	0.1146	1.3251	1.2559
Indirect		**0.0621**	**0.0789**	**0.8510**	0.8575
Direct					**0.7489**
Basic model (Section 5.4.1)	Var.	**0.0513**	0.0671	1.1921	1.1635
Indirect		0.0514	**0.0406**	**0.8796**	**0.8382**
Direct					0.9222

error in terms of Δ75%SUR compared to the indirect mode. However, unlike the AMZ-HD-VJND datasets, the prediction errors of the direct mode also have smaller variance than those of the indirect mode. Additionally, We also observed that the difference between the $\Delta 75\%SUR_{|P-A|}$ and $\Delta 75\%SUR_{|P-E|}$ is relatively higher in AMZ-HDR-VJND datasets than AMZ-HD-VJND datasets. This indicates a larger modeling error for the HDR datasets.

Table 5.14 – Average and variance of prediction error in AMZ-HDR-VJND datasets

Model		ΔSUR		Δ75%SUR	
		|P − A|	|P − E|	|P − A|	|P − E|
Indirect	Mean	0.0589	0.0775	0.8991	1.4838
Direct					**0.9449**
Indirect	Var.	0.0343	0.0365	0.7089	1.0228
Direct					**0.7295**

Conclusion

Compared to the basic parameter-driven model presented in Section 5.4.1, this improved version for SUR/JND prediction includes enhanced feature extraction/selection and regression by incorporating bitstream features and other content features. Our analysis shows that bitstream features have the highest contribution to the prediction of SUR compared to other features. We evaluated two prediction modes: direct and indirect, for predicting the p%SUR. Experimental results demonstrate that our proposed framework outperforms the basic parameter-driven model in Section 5.4.1 for both SUR curve and 75%SUR value predictions.

5.5 Summary

This chapter primarily addresses the objective study of the Satisfied User Ratio (SUR). We first analyzed the resolving power of currently widely used Video Quality Metrics (VQMs) towards SUR. Experimental results show that the investigated VQMs are highly content-dependent and not consistent enough for p%SUR, indicating that the current VQMs are not accurate enough to capture the Just Noticeable Difference (JND). This finding indicates that the widely used VQM struggles to accurately capture both JND and SUR.

Due to the content dependency of VQMs for p%SUR, we proposed a new pipeline to predict p%SUR using VMAF as a proxy. This pipeline consists of two stages: *A)* computing of VQMs and *B)* Regression based on VQMs at the optimal QP/CRF. Experimental results indicate that incorporating quality metrics of compressed video with optimal QP/CRF can significantly reduce the Mean Absolute Error (MAE) of p%SUR predictions using VMAF as a proxy, compared to existing solutions in the literature.

Furthermore, we proposed a parameter-driven model to predict SUR using encoding parameters as proxies. Experimental results show that the proposed parameter-driven model (e.g., 2-p-Logistic) improves the mean SUR prediction error to 0.046, reducing it by 43.64% compared with the baseline, and reduces the mean 75%SUR prediction error from 4.38 QP (baseline) to 2.27 QP. Compared with the SRC-based model, the SRC+PVS-based model slightly improves the mean prediction error of the SUR curve and 75%SUR by 0.0019 and 0.0727 QP, respectively, indicating that the quality degradation features from PVSs are not crucial to SUR prediction.

Finally, we proposed an improved version of the parameter-driven model that includes enhanced feature extraction/selection and regression by incorporating bitstream features and other content features. Our analysis shows that bitstream features have the highest contribution to the prediction of SUR compared to other features. We evaluated two prediction modes: direct and indirect, for predicting p%SUR. Experimental results demonstrate that our proposed framework outperforms the basic parameter-driven model for both the SUR curve and 75%SUR value predictions.

Chapter Contributions

— Analysed the resolving power of different VQM for p%SUR.

— Proposed a pipeline to predict p%SUR using VMAF as a proxy.

— Proposed parameter-driven models to predict SUR using encoding parameters as proxies.

Chapter 6

Application of SUR: Streaming Optimization

Overview

Contents

6.1 Introduction 117
6.2 Bitrate costs for enhanced user satifaction 120
6.2.1 Expanding JND Datasets to other Codecs 123
6.2.2 Bitrate as a Function of Satisfied User Ratio (SUR) . . . 126
6.2.3 Summary 128
6.3 JND aware per-title bitrate ladder optimization 129
6.3.1 JASLA Architecture 130
6.3.2 Evaluation and Results 133
6.3.3 Conclusions 136
6.4 Summary 136

Part of this chapter has been published in research papers [10, 101]

6.1 Introduction

The usage of video streaming platforms such as YouTube, Netflix, Hulu or Amazon Prime Video has become an integral part of our daily lives. In this context, Http Adaptive Streaming (HAS) has become the dominant technique utilized for both live and Video-on-Demand (VoD) streaming applications. HAS relies on Adaptive Bitrate (ABR) Streaming methods which encode video content at multiple bitrate-resolution pairs known as “representations". These different representations are used to construct a so-called bitrate ladder [8], allowing a dynamic adjustment of video quality that takes into account the

available bandwidth of the viewer and the type of device.

So far, the norm has been to utilize a fixed set of representations, such as the HLS bitrate ladder [12] for all video content on a system or platform. However, such a "one-size-fits-all" approach may not be optimal when provisioning a wider range of video content types. As shown in Figure 6.1, for *Dolls_s000*, the cross-over bitrate between 540p and 1080p resolutions happens at approximately 2.0 Mbps, which means at bitrates lower than 2.0 Mbps, 540p resolution outperforms 1080p in terms of VMAF [1]. In comparison, at bitrates higher than 2.0 Mbps, 1080p resolution outperforms 540p. On the other hand, for *RushHour_s000*, 1080p yields higher VMAF over the entire bitrate range, which means 1080p should be selected for the bitrate ladder for the entire bitrate range.

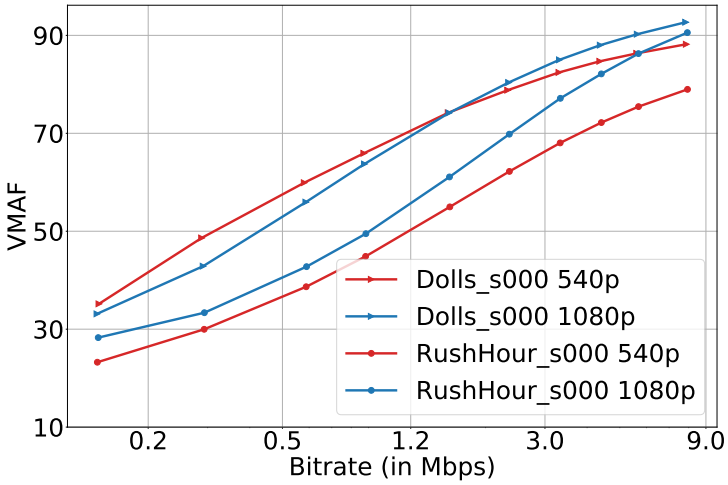


Figure 6.1 – RD curve of 540p and 1080p CBR encodings of *Dolls_s000* and *RushHour_s000* [5] video sequences using x265 HEVC encoder at *slower* preset.

For this reason, *per-title encoding* approaches were introduced, which aim to create an individual, optimized bitrate ladder for each video content in order to achieve higher Quality of Experience (QoE). More specifically, in per-title encoding [29, 8, 6], various encoding parameters (such as frame rate, resolution, *etc.*) are varied and utilized by encoding content clips using all possible combinations of these parameters. Subsequently, an optimized bitrate ladder is constructed by selecting representations from a convex-hull based on the quality measurements of the encoded representations (cf. Figure 6.2). In terms of objective quality metric for the convex-hull construction, VMAF [80] is frequently used, due to its strong correlation with human-perceived quality. Therefore, VMAF is often used to assess the quality of representations and guide the bitrate laddering process [67].

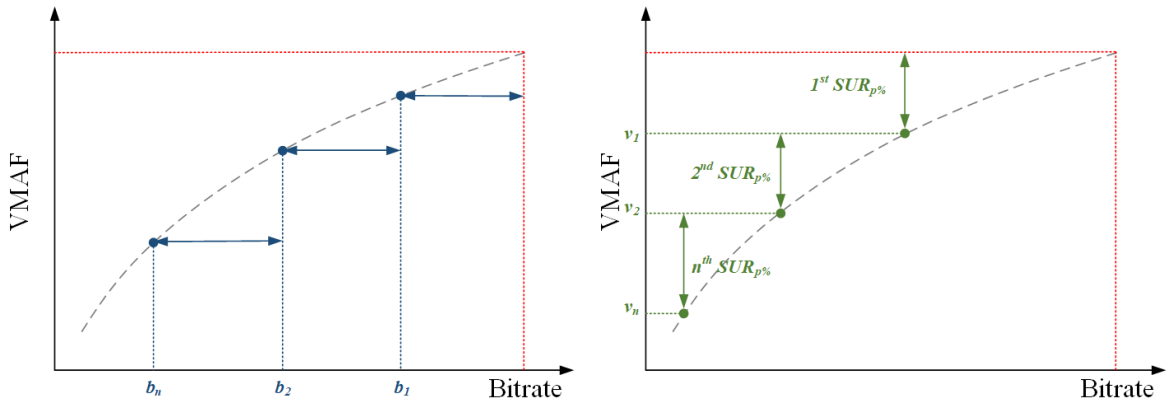
1. https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12, last access: May 30, 2024.

A key step in constructing an optimized bitrate ladder is the careful selection of a subset of representations from the convex-hull. This selection process has to consider a broad range of factors, including available network bandwidth, device capabilities. While some methods focus on selecting representations based on the probability of clients requesting specific bitrate versions [123, 130], other approaches prioritize a selection of representations that minimize perceptual similarity [103, 20, 157]. These methods aim to avoid including representations in the bitrate ladder that feature overly similar perceptual qualities, since such redundancy may lead to inefficient resource utilization. To prevent quality redundancy in the bitrate ladder, the selection process should focus on minimizing perceptual similarity between representations. This approach helps to reduce streaming costs by avoiding the inclusion of perceptually similar quality levels in the bitrate ladder. Figure 6.2 shows an example of selecting bitrate-resolution pairs from the convex-hull. In this example, selection is based on the optimal encoding parameters for each bitrate, by focusing either on bitrate or quality. Figure 6.2a illustrates the selection process based on bitrate, similar to the approach described in [130]. In this method, the most frequently requested set of bitrates, *i.e.,* $\{b_1, b_2, ...b_n\}$, is chosen from within the convex hull to construct the bitrate ladder. Figure 6.2b depicts a quality-based selection process, specifically for VMAF, similar to the approaches outlined in [103, 20]. These methodologies choose a set of quality values, denoted as $\{v_1, v_2, \ldots, v_n\}$ with the goal of having only small, barely noticeable perceptual differences between consecutive representations. These approaches help minimize streaming costs by avoiding the inclusion of perceptually similar quality levels in the bitrate ladder. However, the selection of representations in these works is based on a fixed JND threshold, which may not be optimal for all video content.

In this chapter, we aim to deploy JND and SUR into streaming optimization. We first investigate the bitrate costs associated with varying SUR thresholds to answer the key research question: What is the impact on bitrate when selecting different SUR thresholds? Specifically, how much additional bitrate is required to increase the SUR from 75% to 95%?

To conduct this analysis across various codecs, we first expand the current Just Noticeable Difference (JND) datasets to include additional codecs (Section 6.2.1). We then examine the relationship between SUR thresholds and bitrate requirements for constructing bitrate ladders across these codecs (Section 6.2.2).

In Section 6.3, we introduce a JND-aware bitrate ladder optimization method that leverages the JND values of the video content. This framework, detailed in Section 6.3.1, is designed to construct efficient bitrate ladders. The effectiveness of our approach is



(a) (b)

Figure 6.2 – Optimal representation selection along the convex hull based on (a) bitrate [130] or (b) quality [103] (using VMAF metric).

demonstrated through experimental results presented in Section 6.3.2.

Finally, we conclude the chapter with a summary and suggest directions for future research in Section 6.4.

6.2 Bitrate costs for enhanced user satifaction

In this section, we explore the use of Just Noticeable Differences (JND) for constructing a bitrate ladder with respect to the proportion of satisfied user ratio (SUR). To expand the investigation to various codecs, first, a method is explained that transfers the JND points obtained through subjective testing from one codec (*e.g.,* AVC) to other codecs (*e.g.,* HEVC, VVC). This approach helps avoid the additional costs associated with conducting subjective tests to obtain JND points for a wide range of different codecs. To achieve this objective, we investigate the codec-agnostic nature of various video quality metrics, followed by the transfer of JND between two codecs, taking into account the most suitable codec-agnostic video quality metric. Secondly, we delve into the analysis of the bitrate cost of a given bitrate ladder from a JND perspective, *i.e.,*, as a function of the SUR. Among others, our experimental results demonstrate that increasing SUR leads to an exponential increase in bitrate. For example, to raise the SUR from 75% to 90%, it is necessary to double the video bitrate.

From the sections 3.2.2 of Chapter 3, we know that the JND varies for different observers, and SUR measure the population that could not perceive the difference for a

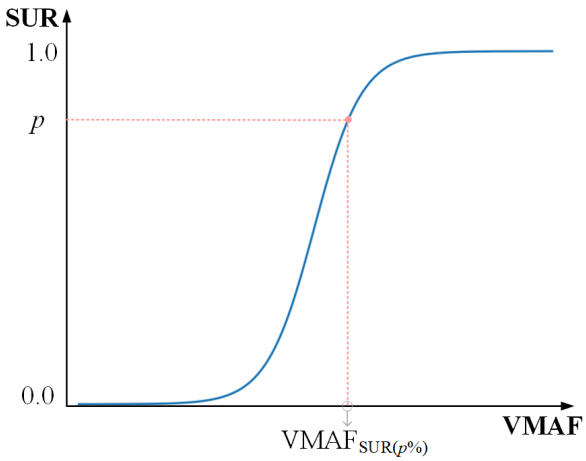


Figure 6.3 – SUR curve example for the VMAF proxy.

certain distortion level. Figure 6.3 presents an example of an SUR curve with VMAF as proxy. Obviously, as the percentage of satisfied individuals p increases (equating a decreasing audience share that is actually capable of perceiving the distortion between the reference and degraded video), the VMAF value shows a corresponding increase. This observation implies that to satisfy a larger number of individuals, the VMAF score of the encoded video should be improved, signifying that the quality of the encoded video should more closely match that of the reference video.

For example, consider $\text{VMAF}_{\text{SUR}(75\%)}$, which signifies that 75% of viewers cannot detect any distortion between the reference video and the encoded video when the VMAF score reaches this level. To cater to a larger audience and meet higher quality standards, where 95% of viewers should not perceive any differences between the reference video and the encoded video, we would refer to the VMAF score as $\text{VMAF}_{\text{SUR}(95\%)}$. In this context, it is important to note that $\text{VMAF}_{\text{SUR}(75\%)}$ would be less than $\text{VMAF}_{\text{SUR}(95\%)}$, indicating a higher level of satisfaction in the latter case.

Figure 6.4 illustrates how various values of p can influence the selection of bitrate-resolution pairs. As the ratio of satisfied individuals, represented by p, increases, the chosen bitrates for the bitrate ladder also increase. In this section, our goals are two-fold:

(i) Our objective is to quantify the relationship between the percentage of satisfied individuals p and the overall bitrate of the selected representations for the bitrate ladder:

$$\sum_{i=1}^{n} b_i = f(p) \tag{6.1}$$

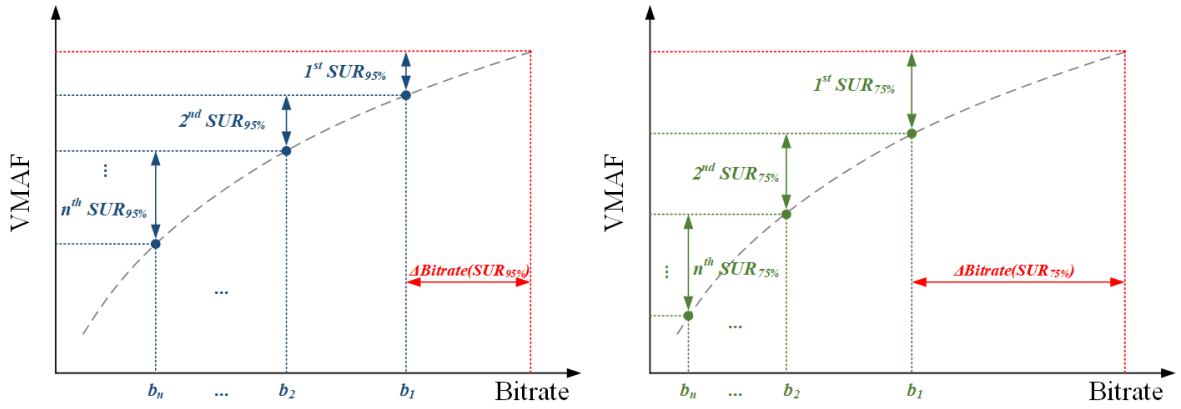


Figure 6.4 – A growing percentage of satisfied users (%p) requires a bitrate ladder with a broader range of bitrate values.

(ii) To accommodate different codecs, we propose a mapping scheme for SUR(p%) across various codecs. To achieve this, we initially assess the codec-dependent nature of different Video Quality Metrics (VQMs) by analyzing a subjective test results. Subsequently, we employ the most codec-agnostic VQM to map SUR(p%) to other codecs. This approach eliminates the necessity for conducting subjective tests for different video codecs. Figure 6.5 illustrates the mapping of the corresponding QP values for SUR(p%) from AVC to VVC.

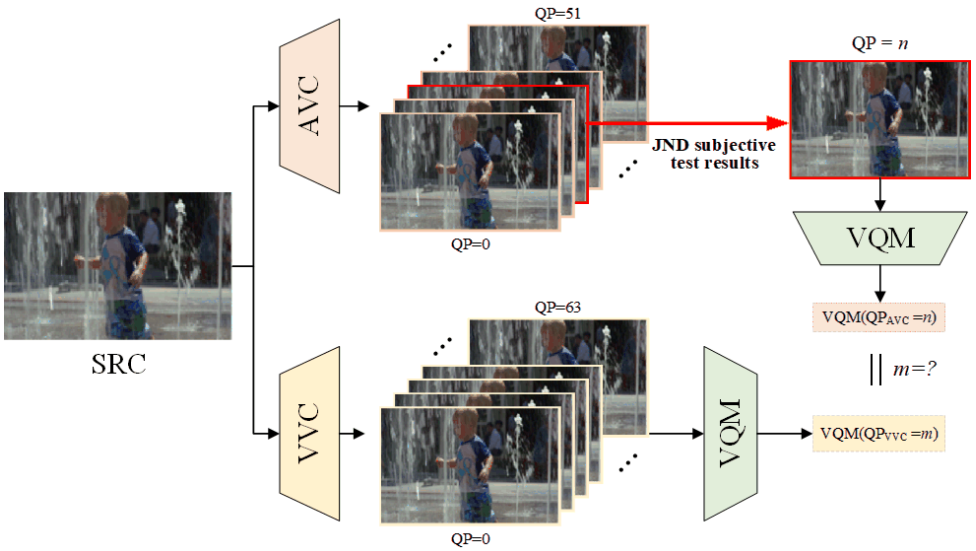


Figure 6.5 – Mapping of the corresponding QP values from AVC to VVC.

6.2.1 Expanding JND Datasets to other Codecs

Carrying out a subjective test is both time-consuming and expensive. This is particularly true for JND subjective tests, as they involve the identification of JNDs within a large set of quality levels [155, 63]. VideoSet [138] is the most extensive JND dataset conducted on the AVC codec to determine the JND values for the QP proxy. The dataset comprises 220 source video sequences with a duration of 5 seconds, featuring frame rates of either 24 fps or 30 fps. These sequences were encoded at different resolutions using the constant quantization parameter (CQP) rate control mode of the AVC, with QPs ranging from 0 to 51. The subjective evaluation of JND of individuals was conducted across multiple universities.

VideoSet currently supports only AVC and to address the need for JND datasets for more advanced codecs like HEVC and VVC, we aim to explore alternatives that do not require an extensive subjective testing process. Our goal is to extend the applicability of JND datasets to other codecs based on VideoSet, originally designed for AVC, while minimizing the need for additional subjective testing efforts. In Figure 6.5, we show the process of mapping the QP from subjective test results of AVC to VVC, all without the need for additional subjective tests. This mapping leverages Video Quality Metrics (VQMs) as proxies, operating under the assumption that widely adopted metrics like VMAF provide consistent scores for the same perceptual quality across different codecs for a given video. In the following section, we substantiate this assumption through a comprehensive analysis of existing cross-codec subjective datasets.

Codec-Agnostic Video Quality Metric (VQM)

We examine the per-content codec-agnostic features of various Video Quality Metrics (VQMs) by leveraging the subjective test results from the Waterloo IVC 4K dataset [81]. The Waterloo IVC 4K dataset includes subjective tests conducted on various video codecs, such as AVC, HEVC, VP9, and AV1. We proceed to compute widely used VQMs, including VMAF (model vmaf_v0.6.1 [2]), PSNR, SSIM, and MS-SSIM, for all the encoded videos. Taking the mean opinion score (MOS) as the ground truth of perceptual quality, *for a VQM to demonstrate codec independence, it should yield the same VQM score for the same MOS across different codecs.*

To quantitatively evaluate the codec-independent performance of various VQMs, we

2. https://github.com/Netflix/vmaf/blob/master/resource/doc/models.md, last access: May 30, 2024.

employ a MOS-VQM regression analysis for each VQM. For a given video encoded with codec $c \in C$ at a VQM set of $\mathbf{Q_c} = [q_c^1, q_c^2, ..., q_c^n]$, we represent the MOS of its encoded versions as $\mathbf{M_c} = [m_c^1, m_c^2, ..., m_c^n]$.

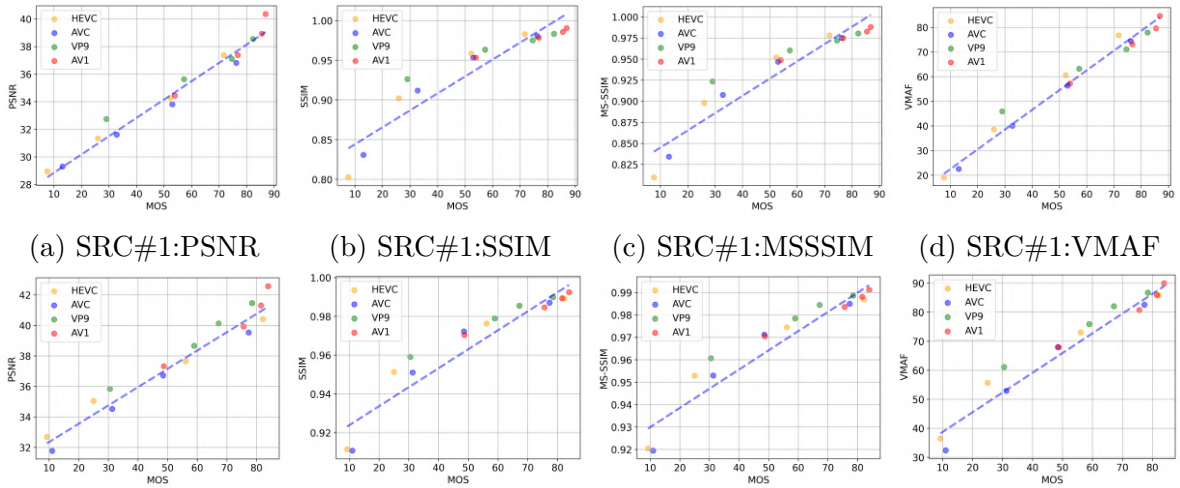
The regression analysis aims to predict MOS values based on VQM scores for videos encoded using various codecs. Consequently, our goal is to identify the optimal VQM and its associated regression model that minimizes the mean absolute error (MAE) across all codecs:

$$\text{MAE} = \frac{1}{N} \sum_{c \in C} \sum_{i=1}^{n} \left| f(q_c^i) - m_c^i \right|, \tag{6.2}$$

where f is obtained from:

$$\min_f \| f(\mathbf{Q_{c_1}}) - \mathbf{M_{c_1}} \|_2^2 . \tag{6.3}$$

This states that for the best fit of the model f, MAE between the predicted MOS values $f(\mathbf{Q}_c)$ and the actual MOS values $\mathbf{M}_c$ over all codecs in set C should be minimal.



(a) SRC#1:PSNR (b) SRC#1:SSIM (c) SRC#1:MSSSIM (d) SRC#1:VMAF

(e) SRC#4:PSNR (f) SRC#4:SSIM (g) SRC#4:MSSSIM (h) SRC#4:VMAF

Figure 6.6 – Relationships between MOS and different VQMs for SRC#1,4 in Waterloo IVC 4K datasets for different codecs.

Given that subjective tests within the VideoSet dataset were performed on AVC only, our analysis selects AVC as c_1 from the codec set $C = \{\text{AVC, HEVC, VP9, AV1}\}$ provided by the Waterloo IVC 4K dataset. For each content in the Waterloo IVC 4K dataset, we fit a regression model as per Eq. (6.3), using the MOS scores associated with AVC encoded

Table 6.1 – MAE and RMSE values between MOS of different codecs and the curve fitted on the AVC codec on Waterloo dataset using linear regression model.

	MAE				RMSE			
Codec	PSNR	SSIM	MS-SSIM	VMAF	PSNR	SSIM	MS-SSIM	VMAF
AVC	**2.86**	6.28	5.73	3.62	**3.08**	6.77	6.14	3.85
HEVC	5.99	7.83	7.38	**5.41**	6.81	9.24	8.86	**6.33**
VP9	10.60	8.50	7.72	**5.86**	11.84	9.60	8.71	**6.88**
AV1	12.98	11.52	10.20	**6.60**	15.37	12.24	10.87	**7.33**
Average	8.11	8.53	7.76	**4.51**	9.28	9.46	8.65	**6.09**

videos and their corresponding VQM scores. Figure 6.6 presents the plots of VQM versus MOS for two different content from the Waterloo IVC 4K dataset, using PSNR, SSIM, MS-SSIM, and VMAF as VQM. The fitted lines are derived from the VQM-MOS data of versions encoded with AVC using a linear regression model.

Table 6.1 summarizes the MAE and root mean sqared error (RMSE) values for the selected VQMs using the linear regression model. Please note that RMSE has been included solely for reference purposes, while MAE has been utilized in the study. MAE measures the absolute difference between the predicted MOS values on the fitted line with AVC MOS-VQM points and the actual human-rated MOS scores. Lower MAE values indicate higher codec-independence of the metric. The results indicate that, using linear regression, VMAF achieves the lowest MAE, at 4.51%, across all codecs, demonstrating its codec independence. This denotes that VMAF consistently estimates the MOS with a marginal error of 4.51% regardless of the codec used.

Generate Cross-Codec JND Datasets

We encoded the 220 source video of VideoSet with HEVC and VVC using the following configurations:

— HEVC: The x265 HEVC video encoder version 3.4, integrated with FFMPEG (libx265), was employed using its default settings (medium preset).

— VVC: The VVenC VVC encoder version 1.9.1[3] [144] was operated using the 'faster' preset and default settings.

VideoSet is based on videos that are encoded with AVC with QP ranging from 0 to 51. For a given content m in VideoSet, assuming that there are N reliable subjects' JND

3. https://github.com/fraunhoferhhi/vvenc, last access: May 30, 2024.

annotations, the JND of a subject “n" is denoted by $j_n^{AVC,m}$. JND of N subjects can be denoted by $J^{AVC,m}$ as

$$J^{\mathrm{AVC},m} = [j_1^{\mathrm{AVC},m}, j_2^{\mathrm{AVC},m}, ..., j_N^{\mathrm{AVC},m}] \tag{6.4}$$

For the same content m for another codec, taking VVC as example, the JND annotations of subject “n" can be computed by Eq. (6.5).

$$j_n^{\mathrm{VVC},m} = \underset{i\in\{1,2,...63\}}{\arg\min} \left|\mathrm{VQM}(\mathrm{QP}_{\mathrm{AVC}}(j_n^{\mathrm{AVC},m})) - \mathrm{VQM}(\mathrm{QP}_{\mathrm{VVC}}(i))\right|, \tag{6.5}$$

where $\mathrm{QP}_c(x)$ is the video encoded with QP of x with codec c. This equation is based on the codec agnostic features of different VQMs which has been validated in Table 6.1.

6.2.2 Bitrate as a Function of Satisfied User Ratio (SUR)

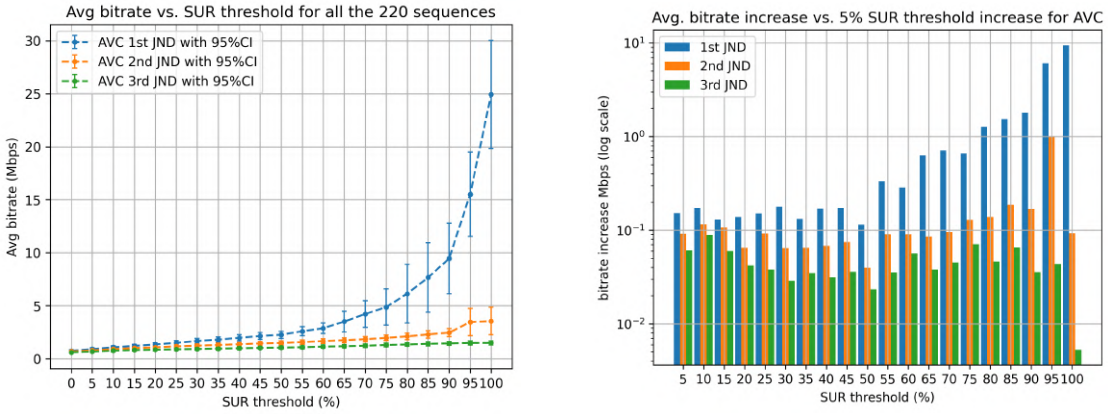
The main objective of the study is to determine the additional bitrate expense required to improve user satisfaction when using various codecs. Accordingly, our initial step involves plotting the average bitrate for videos encoded at QPs corresponding to the SUR(p%) pertinent to different JND levels as a function of the user satisfaction ratio, *i.e.,* p.

Figure 6.7a shows the average bitrates required to satisfy a certain user ratio using the AVC codec, as gathered directly from VideoSet. It has been observed that achieving higher user satisfaction—whereby a greater percentage of users cannot detect any distortion between the reference video and the compressed version—necessitates a greater bitrate. However, the required increase in bitrate to maintain user satisfaction exhibits an exponential trend. For example, to ensure that 75% of users cannot discern any difference at the 1$^{\mathrm{st}}$ JND level, videos need to be encoded at an approximate average of 5 Mbps. If p is increased to 90%, the approximate average required bitrate escalates to 10 Mbps, which represents a significant increase. Figure 6.7b illustrates the increase in bitrate compared to the preceding satisfaction level for each given satisfaction ratio, p. It has also been noted that for the 1$^{\mathrm{st}}$ JND, the increase in bitrate corresponding to an increase in the satisfaction ratio p is greater than that for the 2$^{\mathrm{nd}}$ JND, and the increases for the 2$^{\mathrm{nd}}$ JND are in turn higher than those for the 3$^{\mathrm{rd}}$ JND.

Cross-Codec Comparison on SUR

After generating the new JND datasets for HEVC and VVC as described in Section 6.2.1, we conducted similar analyses to compare them with AVC. To facilitate a clear comparison, we normalized the average bitrate of HEVC and VVC at SUR(p%) by dividing them with the corresponding bitrate of AVC as shown in Eq. (6.6).

$$Bitrate_{normalized}(\mathrm{SUR}^{\mathrm{C}}_{p\%}) = Bitrate(\mathrm{SUR}^{\mathrm{C}}_{p\%}) \Big/ Bitrate(\mathrm{SUR}^{\mathrm{AVC}}_{p\%}) \tag{6.6}$$

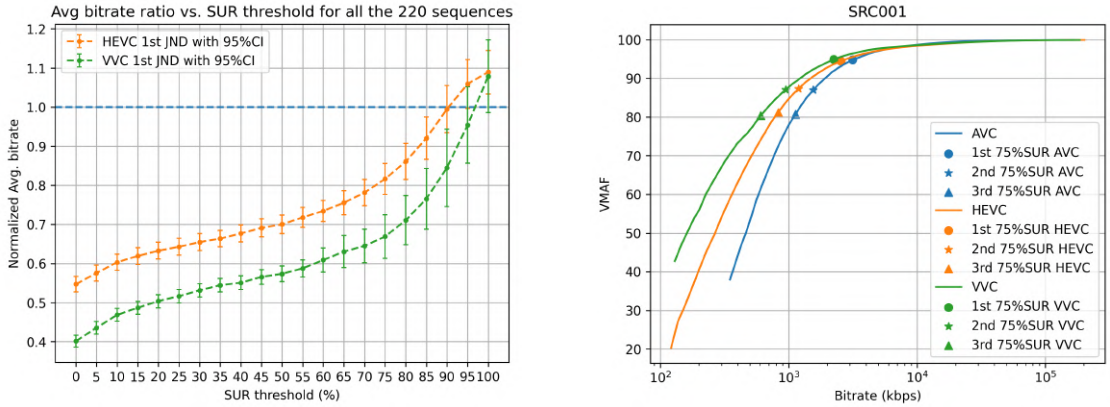


(a) (b)

Figure 6.7 – Relationships between bitrate and SUR threshold for VideoSet AVC 1080p on 1st, 2nd and 3rd JND. (a) Average bitrate for different SUR *thr.* (b) Bitrate increase for every 5% SUR *thr.* increase.

The results are shown in Figure 6.8a. It can be observed that, akin to AVC, the average bitrate required to satisfy a given percentage of users (p%) increases with the p% value. However, the bitrate demand is comparatively lower for HEVC and VVC, as the relative bitrate is less than 1. An increase in relative bitrate can also be noted, suggesting that the efficiency advantages of HEVC and VVC over AVC are more pronounced at lower p% values, with this advantage diminishing as p% rises. For p% equal to 100%, where the average bitrate for AVC is approximately 25 Mbps (refer to Figure 6.7a), the bitrate requirement for HEVC and VVC is greater. This discrepancy may be attributed to the efficiency of AVC in very high bitrates or the specific preset employed for different codecs. Figure 6.8b demonstrates the quantified relationship between VMAF scores and bitrates for the video SRC#1 encoded using different codecs. Additionally, it outlines the associated SUR at a 75% satisfaction level for the 1st, 2nd, and 3rd perceptual JND

thresholds. The data clearly indicates that at identical JND levels, HEVC and VVC achieve the designated quality with lower bitrate requirements than AVC, confirming the superior performance of the former codecs.



(a) (b)

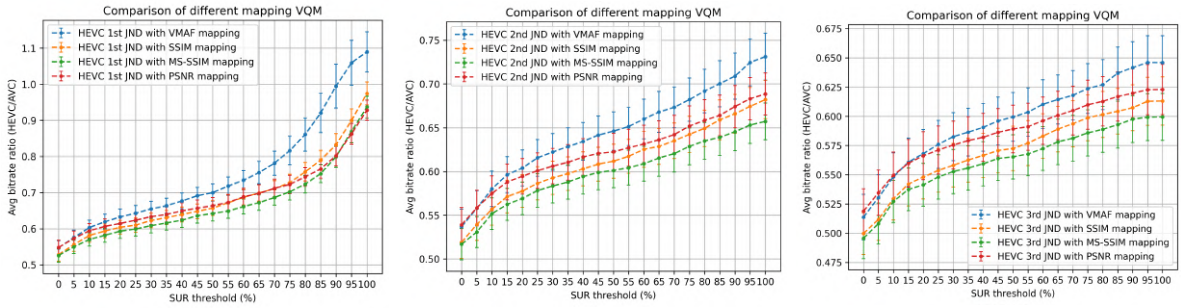
Figure 6.8 – (a) The normalized average bitrate of HEVC and VVC at various SUR(p%)s. (b) VMAF vs bitrate curves for video SRC# 1 encoded with AVC, HEVC, and VVC.

Impact of Video Quality Metric

In Section 6.2.1, we established that VMAF exhibits the highest level of codec-agnostic properties compared to other VQMs. To assess the robustness of our JND mapping methodology across various codecs, we employ multiple VQMs to determine the JND points. This multi-metric approach enables us to understand the extent to which a VQM affects the mapping. Figure 6.9 shows the relative bitrate for the HEVC codec's JND points using different VQMs for mapping. It has been observed that, aside from VMAF, other VQMs yield similar results for the 1st JND, albeit with a significant margin of error compared to VMAF. However, for the 2nd and 3rd JNDs, the similarity between the results diminishes. This highlights the importance of choosing the appropriate VQM for accurate mapping.

6.2.3 Summary

In this section, we conducted an evaluation of the codec-agnostic characteristics of various video quality metrics. Our findings indicate that VMAF exhibits superior robustness



(a) (b) (c)

Figure 6.9 – Comparative mapping results to HEVC videos using different VQMs.

across different codecs, meaning that an equivalent VMAF score for different codecs correlates with a comparable Mean Opinion Score (MOS) for those codecs. Utilizing VMAF, we mapped Just Noticeable Difference (JND) points from the AVC codec within the VideoSet dataset to the HEVC and VVC codecs, based on the premise that equivalent VMAF scores for two codecs correlate to the same perceptual video quality. Subsequently, we assessed the bitrate investment necessary to achieve an increased number of satisfied users across different codecs. Our observations reveal an exponential trend, suggesting that satisfying a larger proportion of users who cannot detect the quality differences between a reference video and its encoded counterpart necessitates a steep increase in bitrate.

6.3 JND aware per-title bitrate ladder optimization

The previous Section 6.2 demonstrates the impact of SUR thresholds on bitrate allocation, providing streaming providers with a clearer understanding of the trade-offs between SUR and bitrate cost. This enables them to determine the optimal SUR threshold for their services. Once the SUR threshold is decided, the next question is how much bitrate and storage cost can be saved by designing the bitrate ladder with the SUR in mind. We call this JND-aware per-title bitrate ladder optimization.

In this section, we propose a JND-aware per-title bitrate ladder optimization framework for adaptive VoD streaming applications, `JASLA`. This framework predicts jointly optimized resolutions and corresponding Constant Rate Factors (CRFs) using spatial and temporal complexity features for a given set of target bitrates for every video content/scene, resulting in an efficient constrained Variable Bitrate encoding. Furthermore, bitrate-resolution pairs that result in distortion below the 1st JND (with a SUR threshold

of 75%) are eliminated, ensuring efficient resource utilization.

6.3.1 JASLA Architecture

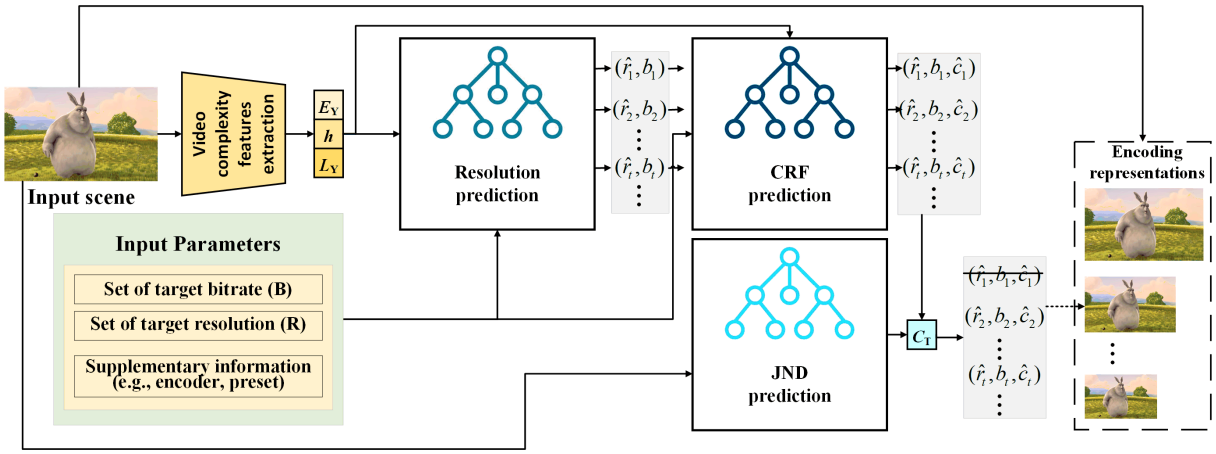


Figure 6.10 – JASLA architecture.

The JASLA architecture is shown in Figure 6.10. The resolution and the corresponding CRF for each bitrate in the bitrate ladder are predicted for every scene using the scene's spatial and temporal complexity features, the set of pre-defined resolutions (R), and the set of pre-defined bitrates (B) for an efficient cVBR steaming. An optimized bitrate ladder for every scene ensures streaming quality with no bitrate fluctuations. R is input to JASLA to confirm that only the resolutions supported by the streaming service provider are selected to generate the optimized bitrate ladder. Next, the bitrate-resolution pairs whose perceptual quality is less than one JND compared to the source video are eliminated. In this way, the number of representations needed for streaming is reduced. The encoding process is carried out only for the predicted bitrate-resolution-CRF pairs for every scene.

JASLA comprises three steps: *(i)* scene complexity features extraction, *(ii)* optimized resolution and CRF prediction, and *(iii)* JND threshold prediction which are described in the following.

Scene Complexity Features Extraction

In video streaming applications, an intuitive method for feature extraction would be to utilize Convolutional Neural Networks (CNNs) [149]. However, such models have several

inherent disadvantages, such as higher training time, inference time, and storage requirements, which are impractical in streaming applications. The popular state-of-the-art video complexity features are Spatial Information (SI) and Temporal Information (TI) [4]. But the correlation of SI and TI features with the encoding output features such as bitrate, encoding time *etc.* are very low, which is insufficient for encoding parameter prediction in streaming applications [97, 103, 134, 104].

In this study, seven DCT-energy-based features [43], the average luma texture energy E_Y, the average gradient of the luma texture energy h, the average luminescence L_Y, the average chroma texture energy E_U and E_V (for U and V planes) and the average chrominance L_U and L_V (for U and V planes), which are extracted using VCA [5] open-source video complexity analyzer [105, 104] are used as the spatial and temporal complexity measures [98, 99] of every scene.

Optimized Resolution and CRF Prediction

For each scene, the optimized resolution for a given target bitrate is predicted using the scene's spatial and temporal features, the set of supported resolutions (R), and the set of target bitrates (B). To determine the bitrate-resolution pairs of the bitrate ladder, VMAF is predicted for each target bitrate (b_t) in the set B for all resolutions $\tilde{r}$ in R, denoted as $v_{\tilde{r},b_t}$. From the predicted VMAF values, the resolution which yields the maximum VMAF value is chosen as the optimized resolution for the target bitrate. Random Forest (RF) models are trained to predict VMAF for every resolution supported by the streaming service provider. This ensures *scalability* of design, where there is no requirement to retrain the entire network to add a new resolution to the framework.

Using the E_Y, h, L_Y features, optimized CRF $\hat{c}_t$ is estimated for every ($\hat{r}_t$, b_t) representation of the bitrate ladder for cVBR encoding. Prediction models are trained for each resolution $\tilde{r}$ in R, which determines $\hat{c}_t$ based on E_Y, h, L_Y and $log(b_t)$ for every scene. The minimum and maximum CRF (c_{min} and c_{max}, respectively) are chosen based on the target codec. For example, x265[6] supports a CRF range between 0 and 51. The prediction algorithm for the bitrate ladder is shown in Algorithm 4.

4. https://github.com/VQEG/siti-tools, last access: May 30, 2024.
5. https://vca.itec.aau.at, last access: May 30, 2024.

Algorithm 4 Optimized resolution and CRF prediction

Inputs:
  R : set of all resolutions $\tilde{r}_m \ \forall \ m \in [1, M]$
  M : number of resolutions in R
  B : set of all bitrates $b_t \ \forall \ t \in [1, N]$
  N : number of bitrates in B
  E_Y, h, L_Y : average scene complexity
Output: $(\hat{r}, b, \hat{c})$ pairs of the bitrate ladder
for $t \in [1, N]$ **do**
    for $m \in [1, M]$ **do**
        Determine $v_{\tilde{r}_m, b_t}$ with $[E_Y, h, L_Y, log(b_t)]$, using the model trained for $\tilde{r}_m$.
    end for
    $\hat{r}_t = \arg\max_{\tilde{r}_m \in R}(v_{\tilde{r}, b_t})$
    Determine $\hat{c}_t$ with $[E_Y, h, L_Y, log(b_t)]$, using the model trained for $\hat{r}_t$.
  $(\hat{r}_t, b_t, \hat{c}_t)$ is the $(t)^{th}$ point of the bitrate ladder
end for

JND Threshold Prediction

We adapted the SUR prediction framework described in Section 5.4.2 of Chapter 5 to predict the CRF value (c_T) where 75% of observers cannot perceive any distortion compared to the source video. The original model, shown in Figure 5.12, utilizes three types of features: (i) masking effect features, (ii) bitstream features, and (iii) content features. However, computing the masking effect features is very time-consuming.

To address this, we replaced the masking effect features with scene complexity features: E_Y, h, L_Y. These features are already extracted for resolution and CRF prediction, as shown in Figure 6.10. The prediction results after this modification are presented in Table 6.2. We observed that the prediction error increased by only 0.2 compared to the original model. Despite this slight increase in error, the revised model is significantly more efficient for bitrate ladder optimization: for the same 10s video of 1080p, using a computer with an Intel(R) Xeon(R) CPU E7-8870 v4 @ 2.10GHz, extracting the masking effect features takes 506 minutes, while extracting the scene complexity features takes just 0.9 seconds. It's important to note that the implementation of the masking effect features is not optimized, and the extraction time could be reduced through parallel computation.

Representation elimination: c_T is used to eliminate perceptually redundant representations from the bitrate ladder as shown in Algorithm 5. There shall be only one representation in the bitrate ladder where the selected optimized resolution is the maximum supported resolution (r_{max}), and the predicted optimized CRF is lower than c_T. Other

Table 6.2 – Mean and Variance of the JND threshold prediction error for complexity reduced model on AMZ-HD-VJND dataset.

Model	Mean ($\Delta SUR_{\lvert P-A \rvert}$)	Var ($\Delta SUR_{\lvert P-A \rvert}$)
Original model	0.7498	0.9222
Complexity reduced model	0.9491	1.0888

higher bitrate representations are eliminated.

Algorithm 5 Representation elimination

Inputs:
N : number of bitrates in B
$(\hat{r}, b, \hat{c})$ pairs of the bitrate ladder
c_T : JND threshold CRF
r_{max} : maximum resolution in R
Output: $(\hat{r}, b, \hat{c})$ pairs for encoding $t = 1, flag = 0$
while $t \leq N$ **do**
if $\hat{r}_t == r_{max}$ and $\hat{c}_t < c_T$ **then**
$flag++$
end if
if flag > 1 **then**
Eliminate $(\hat{r}_t, b_t, \hat{c}_t)$ from the ladder.
end if
$t++$
end while

6.3.2 Evaluation and Results

Test Methodology

In this study, four hundred video sequences (*i.e.,* 80% of all sequences) from the Video Complexity Dataset [5] are used as the training dataset, and the remaining (20%) is used as the test dataset. The video sequences are encoded at 30fps using x265 [6] v3.5 with the *slower* preset. The bitrate-ladder specified in Apple HLS authoring specifications [7] are considered in the evaluation, *i.e.,* R= {360p, 432p, 540p, 720p, 1080p} and B = {145, 300, 600, 900, 1600, 2400, 3400, 4500, 5800, 8100}. E_Y, h and L_Y features are extracted using VCA[5] v1.5 open-source video complexity analyzer [104] run in eight CPU threads

6. https://videolan.org/developers/x265.html, last access: May 30, 2024.
7. https://developer.apple.com/documentation/http-live-streaming/hls-authoring-specification-for-apple-devices, last access: May 30, 2024.

using x86 SIMD optimization [132]. Hyperparameter tuning is performed to obtain a balance between the model size and performance for VMAF and CRF prediction models, which results in the following parameters [8] for VMAF and CRF prediction models: $min\_samples\_leaf = 1$, $min\_samples\_split = 2$, $n\_estimators = 100$, and $max\_depth = 14$. Furthermore, the bitstream features are extracted from the CRF=5 encoded bitstream for each scene. $Q \times Q$ is set as 64×64 to determine GLCM features. The JND prediction model is trained on our collected AMZ-HD-VJND datasets 3. The kernel of SVR is the Radial basis function [9] with the parameters $\epsilon = 0.0001$ and regularization parameter $C = 0.1$ determined by a greedy hyperparameter search.

The following metrics are considered during the evaluation: *(i)* quality in terms of PSNR and VMAF[1], *(ii)* bitrate, and *(iii)* encoding time. Since the content is assumed to be displayed at Full HD (1080p) resolution [29], the encoded content is scaled to 1080p resolution, and VMAF and PSNR are calculated. Bjøntegaard delta rates [21] BDR_P and BDR_V refer to the average increase in bitrate of the representations compared with that of the fixed bitrate ladder encoding to maintain the same PSNR and VMAF, respectively. BD-PSNR and BD-VMAF refer to the average increase in PSNR and VMAF, respectively, at the same bitrate compared with the reference bitrate ladder encoding scheme. The relative difference in the storage space required to store all representations (ΔS) is also evaluated as:

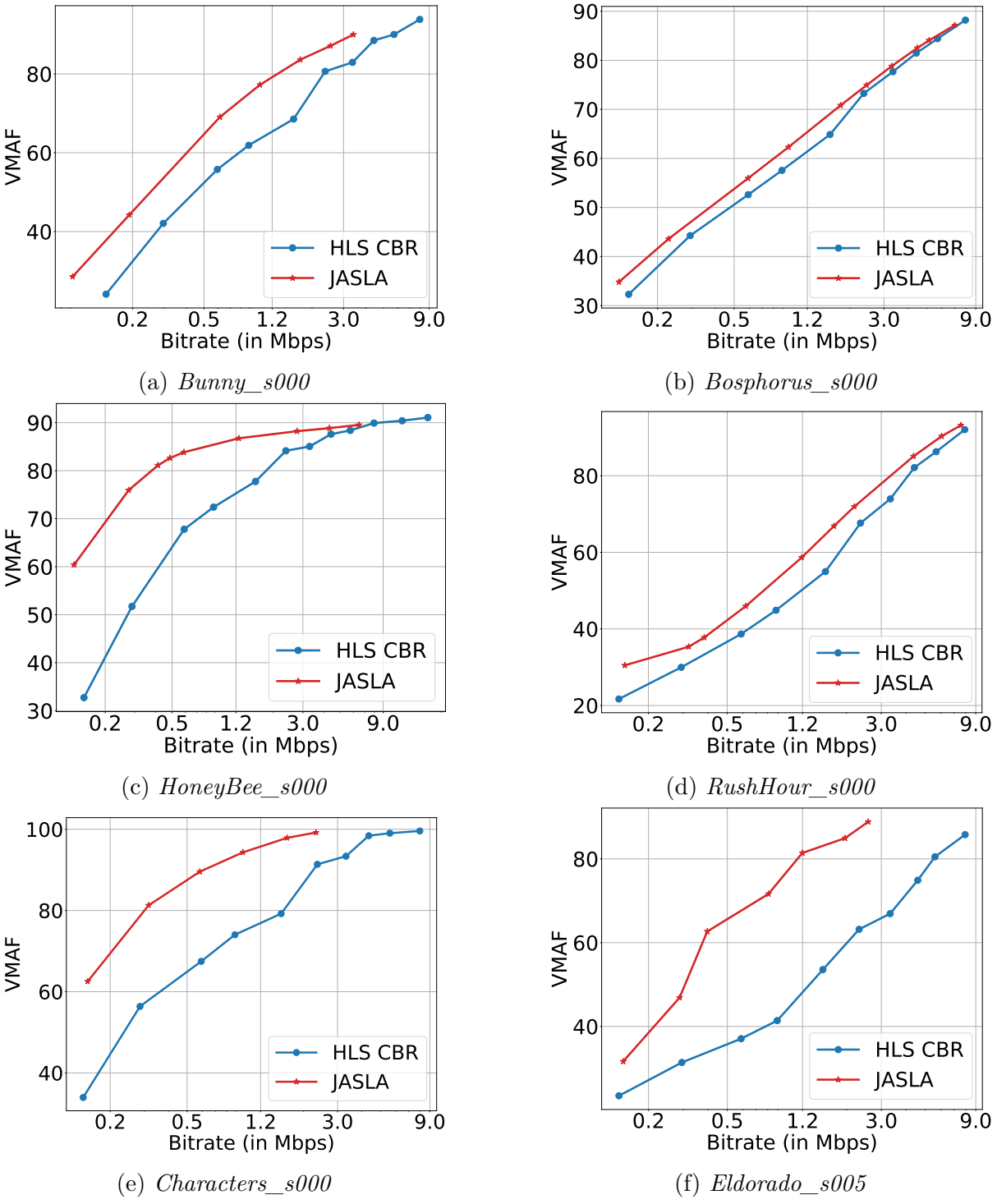
$$\Delta S = \frac{\sum b_{opt}}{\sum b_{ref}} - 1 \tag{6.7}$$

where $\sum b_{ref}$ and $\sum b_{opt}$ represent the sum of bitrates of all representations in the reference bitrate ladder encoding and `JASLA` encoding, respectively.

Experimental Results

The performance of the VMAF, CRF, and JND threshold prediction models is investigated in the first experiment. The average R^2 score of the VMAF and CRF prediction models are estimated as 0.93 and 0.97, respectively. Hence, a strong positive correlation exists between the predicted and ground truth values. The average MAE of the prediction models is estimated as 3.25 and 1.86, respectively. The MAE of the JND threshold prediction model is observed to be 0.94, which shows that `JASLA` works with sufficient

8. https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees, last access: May 30, 2024.
9. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html, last access: May 30, 2024.



(a) *Bunny_s000*

(b) *Bosphorus_s000*

(c) *HoneyBee_s000*

(d) *RushHour_s000*

(e) *Characters_s000*

(f) *Eldorado_s005*

Figure 6.11 – Comparison of RD curves of representative scenes using HLS CBR encoding (blue line), `JASLA` encoding (red line).

prediction accuracy.

The second experiment analyzes the runtime complexity of `JASLA`. `JASLA` predicts resolution and CRF at a rate of 300 frames per second, *i.e.,* 0.4s per video segment. Compared to [154], the JND prediction runtime in `JASLA` is decreased by 97.24%.

The third experiment analyzes the bitrate saving and storage reduction results of `JASLA` compared to the HLS CBR encoding. Using `JASLA` encoding, BDR_P, BDR_V, and ΔS are observed as -34.42%, -42.67% and -54.34%, respectively, compared to the HLS CBR encoding. Moreover, `JASLA` encoding yields an average BD-PSNR and BD-VMAF of 2.90 dB and 9.51, respectively. Figure 6.11 shows the RD curves of eight representative video sequences (scenes) with HLS CBR encoding and `JASLA` encoding. The representative scenes exhibit a variety of spatial and temporal complexities (in terms of E_Y, h, and L_Y). `JASLA` yields the highest VMAF at the same target bitrates for all scenes. Moreover, the perceptually lossless representations are eliminated from the bitrate ladder.

6.3.3 Conclusions

we proposes a JND-aware per-scene bitrate ladder prediction scheme (`JASLA`) for adaptive video-on-demand streaming applications. `JASLA` predicts the optimized resolution and corresponding CRF for given target bitrates for every video scene based on content-aware spatial and temporal complexity features. A JND threshold prediction scheme is proposed, eliminating representations that yield distortion lower than one JND from the bitrate ladder. The performance of `JASLA` is analyzed using the x265 open-source HEVC encoder against a standard HLS bitrate ladder with the maximum resolution of Full HD (1080p). It is observed that, on average, streaming using `JASLA` requires 34.42% and 42.67% fewer bits to maintain the same PSNR and VMAF, respectively, compared to the reference HLS bitrate ladder, along with a 54.34% cumulative decrease in the storage space needed to store representations.

6.4 Summary

In this chapter, we explore the application of VW-JND in video streaming. We begin with a study on the impact of the SUR threshold on bitrate allocation in adaptive streaming. The experimental results reveal an exponential relationship between the SUR threshold and bitrate cost. For instance, increasing the SUR from 75% to 90% requires

doubling the video bitrate. This study provides streaming providers with a clearer understanding of the trade-offs between SUR and bitrate cost, enabling them to determine the optimal SUR threshold for their services. We also examine the codec-dependent nature of various VQMs and propose a mapping approach to expand VW-JND datasets across codecs.

Next, we propose a JND-aware per-title bitrate ladder optimization framework for adaptive VoD streaming applications, `JASLA`. By leveraging the VW-JND prediction framework, we eliminate perceptually lossless representations from the bitrate ladder. Experimental results show that `JASLA` requires 34.42% and 42.67% fewer bits to maintain the same PSNR and VMAF, respectively, compared to the reference HLS bitrate ladder. Additionally, there is a 54.34% cumulative decrease in the storage space needed to store the representations.

The proposed `JASLA` framework currently focuses solely on the first JND threshold for determining the maximum resolution of the bitrate ladder. Eliminating perceptually lossless representations can save significant resources, given their high resource demands. However, there is potential for further optimization. Future work could explore integrating the second and third JND thresholds, offering opportunities to refine the bitrate ladder and enhance resource efficiency even further.

Chapter Contributions

- Explored the impact of SUR thresholds on bitrate allocation in adaptive streaming, revealing an exponential relationship.
- Evaluated the codec-dependent nature of various VQMs and devised a mapping approach to expand VW-JND datasets across codecs.
- Introduced `JASLA`, a JND-aware bitrate ladder optimization framework for adaptive VoD streaming.

Chapter 7

Conclusion

Overview

Contents

7.1 Summary of contributions 139
7.2 Limitations and Perspectives 141

7.1 Summary of contributions

In-the-wild subjective test system: (1) Development of AtHome Subjective Test System: We introduced a novel AtHome subjective test system designed for high-end video quality assessments. (2) Comparison of Viewing Environments: Utilizing the AtHome system, we analyzed the impact of viewing environments, specifically comparing "AtHome" and "InLab" settings. Experimental results demonstrate that the opinion scores obtained from both environments are not significantly different. However, the confidence intervals (CI) of the opinion scores are larger in the AtHome environment, indicating greater variability in responses. (3) Impact of Display on Opinion Scores: The AtHome system's flexibility with different displays allowed us to analyze the impact of various display technologies on opinion scores. Experimental results indicate that the variety and advanced functionalities of display ecosystems significantly affect the Quality of Experience (QoE) for end users.

Subjective study of VW-JND: (1) Generalization of SUR Definition: We compared related work on VW-JND and extended the definition of the Satisfied User Ratio (SUR) of JND to include different proxies such as Video Quality Metrics (VQM) and encoding parameters. (2) Benchmarking JND Search Methodologies: Through simulation, we benchmarked various JND search methodologies and demonstrated that the Relaxed Binary Search (RBS) is more efficient in terms of experiment time compared to Simple

Staircase and Quest+ methods. (3) Pre-processing Method for JND Candidate Playlist (JCP): We proposed a pre-processing method for the JND Candidate Playlist (JCP) that reduces the duration of subjective tests by over 9%, enhancing the efficiency of data collection. (4) Collection of HD and HDR VW-JND Datasets: Using our AtHome subjective test system, we collected VW-JND datasets for HD and HDR content for high-end video quality assessment.

Subjective Data Analyses: (1) ZREC Method for Opinion Score Recovery: We proposed ZREC, a robust method to recover mean and percentile opinion scores. Experimental results show that parameters such as subject bias and inconsistency, content ambiguity estimated by ZREC correlate highly with more complex solver-based methods and standards. Additionally, ZREC recovers Mean Opinion Scores (MOS) with smaller confidence intervals than current state-of-the-art methods. (2) Impact of ZREC on SUR Prediction: Experimental results indicate that using recovered percentile opinion scores of ZREC as ground truth during training improves the performance of SUR prediction models. (3) Estimating Uncertainty: we introduced mathematical methods to estimate the uncertainty of both $p\%\text{SUR}_{\text{emp}}$ and the SUR curve. This is crucial for understanding subjective data but has been largely ignored in the literature. (4) Longitudinal Analysis with AtHome System: Our AtHome subjective test system allows participants to conduct tests over an extended period, enabling cross-campaign and intra-campaign analysis. Experimental results show that observer bias and inconsistency remain relatively stable over time.

Objective study of SUR: (1) Resolving Power of VQM: We analyzed the resolving power of VQMs for SUR and found that current widely used VQMs are highly content-dependent for p%SUR. This presents a new challenge for VQM development: a good VQM should not only have a high correlation with Mean Opinion Scores (MOS) but also be consistent for p%SUR. (2) SUR Prediction Using VMAF as proxy: We proposed a new pipeline to predict p%SUR using VMAF as a proxy. Experimental results show that the proposed method can predict p%SUR with a Mean Absolute Error (MAE) of 1.67 on VideoSet and 0.29 on AMZ-HDR-VJND datasets. (3) Parameter-Driven Models for SUR Prediction: We also proposed a parameter-driven model to predict SUR using encoding parameters as a proxy. The parameter-driven model (e.g., 2-p-Logistic) improves the mean SUR prediction error to 0.046, reducing it by 43.64% compared with the baseline, and reduces the mean 75%SUR prediction error from 4.38 QP (baseline) to 2.27 QP, with only the SRC as input without extensive recompression (4) Improved Parameter-Driven

Models: We further improved the parameter-driven model by including enhanced feature extraction/selection and regression by incorporating bitstream features and other content features. Our analysis shows that bitstream features have the highest contribution to the prediction of SUR compared to other features. We evaluated two prediction modes: direct and indirect, for predicting p%SUR. Experimental results demonstrate that our proposed framework outperforms the basic parameter-driven model for both the SUR curve and 75%SUR value predictions.

Application of SUR in Streaming: (1) Bitrate Costs for SUR: Experimental results show that increasing the SUR leads to an exponential increase in bitrate across different codecs. (2) Examination of Codec-Agnostic Features of VQM: We analyze the per-content codec-agnostic features of different VQMs to extend the VW-JND datasets to other codecs. (3) Applying SUR to Streaming Systems: We propose a JND-aware per-title bitrate ladder optimization framework for VoD streaming applications. Experimental results indicate that this framework can save up to 54.34% of storage space and requires 34.42% and 42.67% fewer bits to maintain the same PSNR and VMAF score, respectively, compared to the HLS bitrate ladder.

7.2 Limitations and Perspectives

This thesis primarily focuses on the first JND threshold for the maximum resolution in bitrate ladders. While eliminating perceptually lossless representations can significantly save resources, the scope is restricted to the first JND. Future work could explore the 2nd and 3rd JNDs of different resolutions. One major challenge in conducting subjective JND research across resolutions is how to order the JND candidate playlist. All JND search methods mentioned in Section 3.3 are based on the premise that the JND candidate playlist is ordered by perceptual quality. For a single resolution, we can order the JCP using encoding parameters, but for cross-resolution research, we need to find a way to order the JCP by perceptual quality, which remains an open question in the field of video quality assessment.

The AtHome subjective test system provides a more ecological testing environment compared to traditional InLab settings. Even though we provide instructions on the ambient light and viewing distance and visited participants' homes to measure the ambient light, the current AtHome subjective test system doesn't allow us to measure and verify the actual test conditions, especially the ambient light and viewing distance, through-

out the entire test campaigns. Future work could explore the possibility of supervising ambient light and viewing distance during the test campaigns.

While the AtHome system allows for long-term testing and helps us conduct cross-campaign and intra-campaign analysis, the results are based on a relatively small sample size. Conducting larger-scale and longer-duration longitudinal studies with the AtHome system can provide deeper insights into observers' behavior and consistency over time. This would help in understanding the long-term stability of subjective assessments.

We compared various JND search methods and proposed a pre-processing method for the JCP, which effectively reduces the duration of subjective tests. However, the JND search process remains time-consuming. Additionally, our uncertainty analysis of the SUR curve obtained from the JND subjective datasets reveals that current JND search methods introduce considerable uncertainty to the SUR curve. Therefore, future work should explore more efficient methodologies for JND search to reduce both the JND search time and the uncertainty of the SUR curve.

Although efforts were made to extend VW-JND datasets to other codecs in Section 6.2.1, the analysis remains somewhat codec-dependent. Developing truly codec-agnostic VQMs that maintain high accuracy and precision across different codecs will be crucial. This would involve more extensive testing and validation across a wider range of video content and codecs.

In this thesis, the application of the SUR into bitrate ladder optimization for HAS is limited to the VoD streaming use case due to the complexity of the JND prediction. Future work could explore the application of SUR in other streaming use cases, such as live streaming, by reducing the complexity of the JND prediction model.

List of Abbreviations

The list of the abbreviations used in this thesis in alphabetical order.

— ACR: Absolute Category Rating

— ABR: Adaptive Bitrate Streaming

— ANOVA: Analysis of Variance

— CBR: Constant Bitrate

— CI: Confidence Interval

— CDF: Cumulative Distribution Function

— CCDF: Complementary Cumulative Distribution Function

— CNN: Convolutional Neural Network

— CRF: Constant Rate Factor

— CLT: Central Limit Theorem

— COV: Coefficient of Variation

— cVBR: Constrained Variable Bitrate

— DCR: Degradation Category Rating

— DMOS: Differential Mean Opinion Score

— EBA: Eliminated by Aspects

— FoV: Field of View

— GLCM: Gray Level Co-occurrence Matrix

— HAS: Http Adaptive Streaming

— HDR: High Dynamic Range

— HD: High Definition

— HMD: Head Mounted Display

— HVS: Human Visual System

— ITU: International Telecommunication Union

— i.i.d: Independent and Identically Distributed

— JCP: JND Candidate Playlist

— JND: Just Noticeable Difference

— JPEG: Joint Photographic Experts Group

— K-S: Kolmogorov-Smirnov

— MAE: Mean Absolute Error

— MaxCLL: Maximum Content Light Level

— MaxFALL: Maximum Frame Average Light Level

— MOS: Mean Opinion Score

— MLE: Maximum Likelihood Estimation

— PDF: Probability Density Function

— PEST: Parameter Estimation by Sequential Testing

— PMF: Probability Mass Function

— PVS: Processed Video Sequence

— PW-JND: Picture-Wise Just Noticeable Difference

— POS: Percentile Opinion Score

— PLCC: Pearson Linear Correlation Coefficient

— QF: Quality Factor

— QoE: Quality of Experience

— QP: Quantization Parameter

— RBS: Relaxed Binary Search

— SI: Spatial Information

— SDR: Standard Dynamic Range

— SOS: Standard deviation of the Opinion Scores

— SRC: Source

— SRCC: Spearman Rank Correlation Coefficient

— SUR: Satisfied User Ratio

— SVR: Support Vector Regression

— TI: Temporal Information

— UHD: Ultra High Definition

— VMAF: Video Multimethod Assessment Fusion

— VOD: Video On Demand

— VW-JND: Video-Wise Just Noticeable Difference

— WCG: Wide Color Gamut

ANNEXES

A System bias for SUR curve estimation?

In the literature [138, 135, 151, 152], the analytical SUR curve is defined as the Complementary Cumulative Distribution Function (CCDF) of the distribution of individual observers' Just Noticeable Difference (JND) annotations obtained from subjective tests. This definition aligns with Case 1 of our extended SUR definition for any proxy as discussed in Section 3.2.2 of Chapter 3.

However, a recently published paper [63] highlights a system bias in the SUR curve when following this definition. The bias arises from the assumption that the JND of observer i for a given content m is not a constant value but a random variable $J_{i,m}$ that follows a Gaussian distribution with mean μ_i and standard deviation σ.

The SUR curve is subsequently modified from Eq. (7.1) to Eq. (7.2). For a detailed explanation, please refer to [63].

$$\mathrm{SUR}_{analy}(x) = 1 - \Phi(\frac{x - \bar{\mu}}{\sigma_0}) \tag{7.1}$$

$$\mathrm{SUR}_{analy\_unbiased}(x) = 1 - \Phi(\frac{x - \bar{\mu}}{\sqrt{{\sigma_0}^2 + \sigma^2}}) \tag{7.2}$$

Where $\bar{\mu}$ and σ_0 are the mean and standard deviation of the individual JND thresholds μ_i across different observers, respectively. σ represents the standard deviation of the individual JND threshold for observer i. As shown in Figure A.1, the probability density function (PDF) of the unbiased SUR curve is more spread out than the original one. Consequently, the SUR curve obtained using Eq. (7.1) overestimates the SUR value when $x < \bar{\mu}$ and underestimates the SUR value when $x > \bar{\mu}$.

This is based on the fact that if we repeat the exact same JND subjective test for content m multiple times for the same observer, the JND threshold for this person would not be exactly the same every time. However, Eq. (7.1) assumes that the random sample taken from the observer at a given time is the mean value μ_i of this observer. In other words, [63] emphasized that the JND threshold for each observer is not a constant value

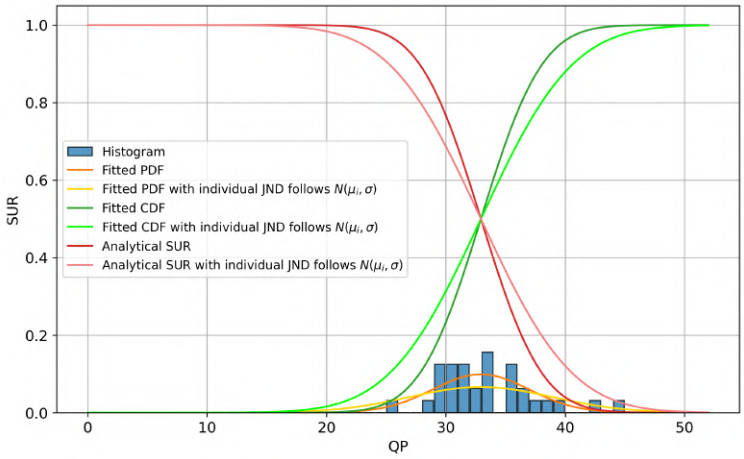


Figure A.1 – Illustration of the system bias for SUR curve estimation of SRC12 in VideoSet [138] according to [63]

but a random variable. However, in SOTA, the individual JND threshold is not modeled as a random variable but rather as a constant value, typically obtained via binary search [138]. This is a general confusion in the entire field of multimedia JND research.

However, the assumption that the standard deviation of all observers' JND random variables is the same in Eq. (7.2) is not the general case, because different observers can be more or less consistent and therefore have different standard deviations. But this assumption still reveals the inaccuracy of the current SUR modeling.

In this thesis, we address this problem by modeling the uncertainty of the SUR in Section 4.3 of Chapter 4. As shown in Figure A.2, the system bias illustrated in Figure A.1 is accounted for within the 95% confidence interval (CI) of the SUR curve.

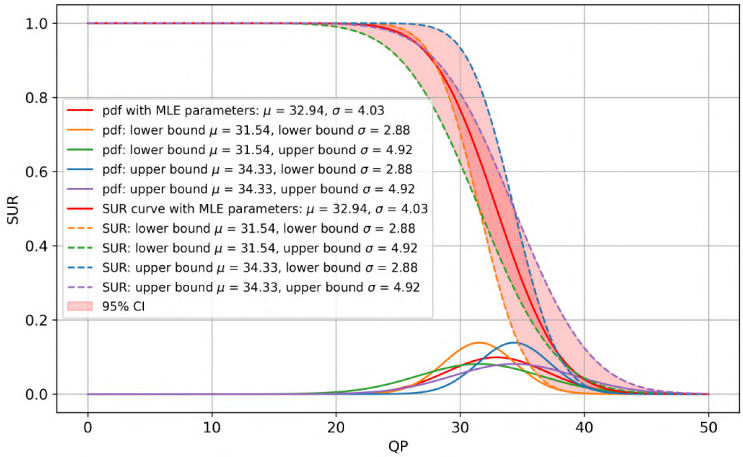
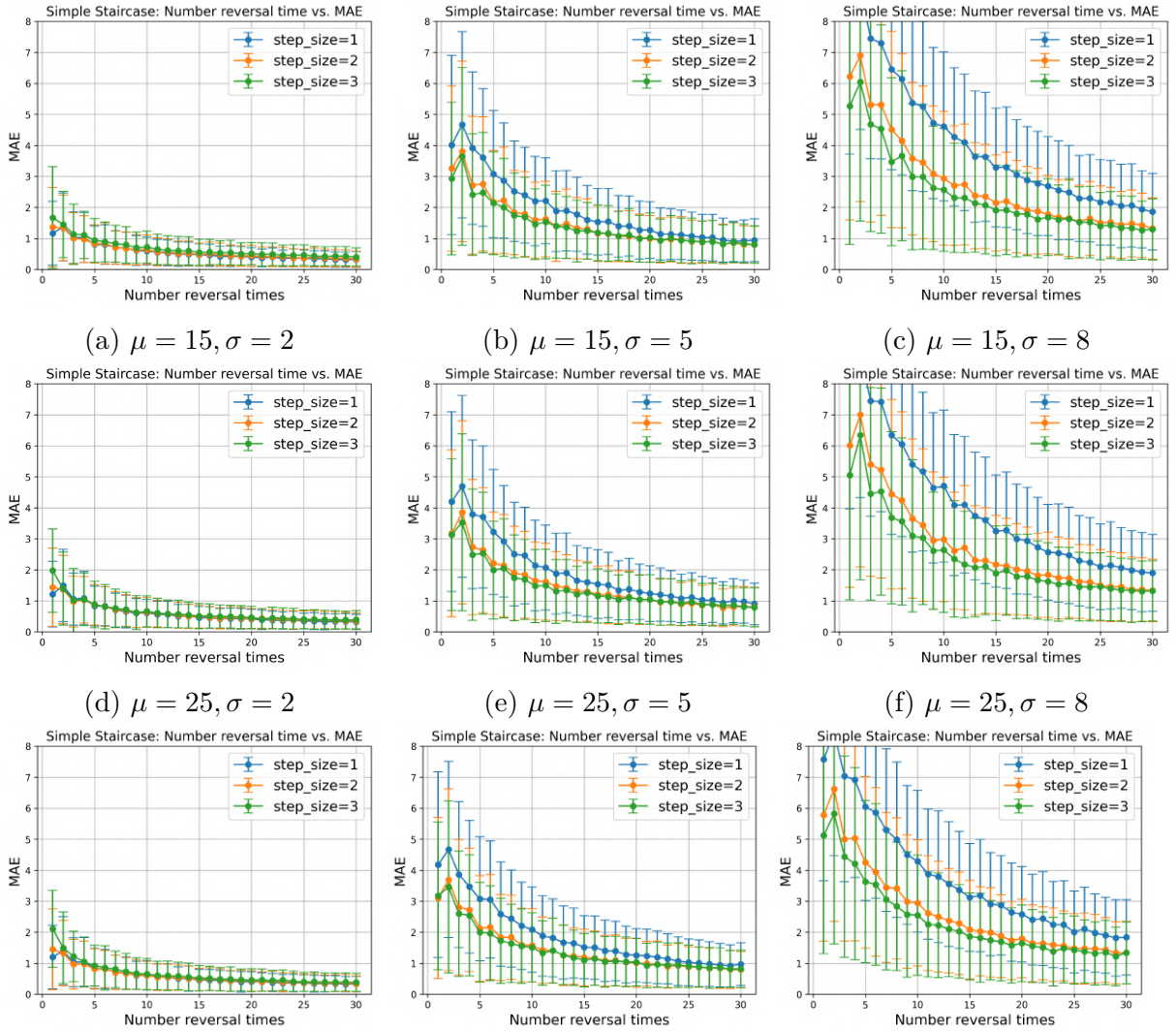


Figure A.2 – MLE estimated SUR curve and 95%CI for SRC12 in VideoSet [138] 1080p with Gaussian assumption. More detials please refer to Section 4.3 of Chapter 4.

B Simple Staircase simulation



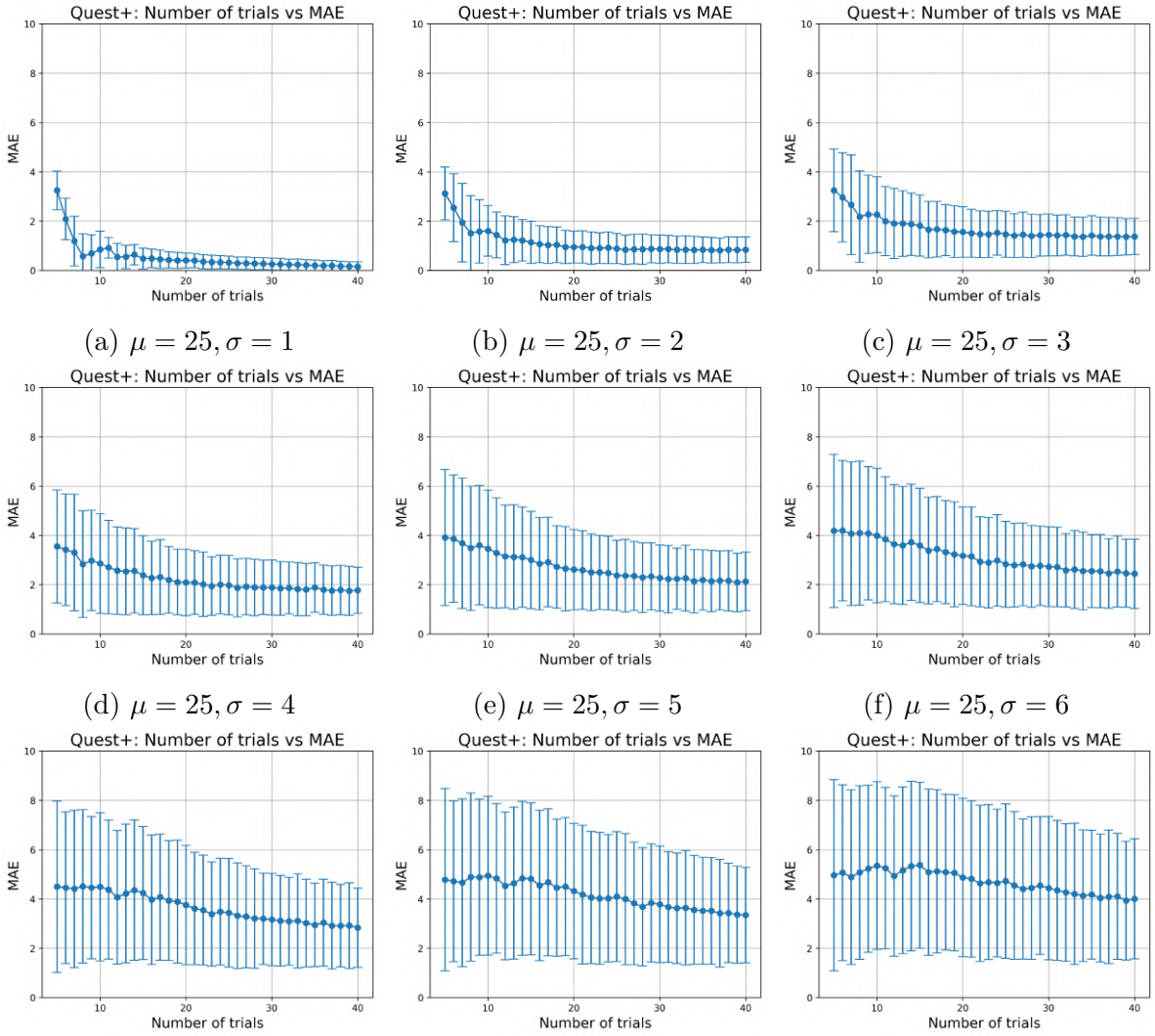
(a) $\mu = 15, \sigma = 2$ (b) $\mu = 15, \sigma = 5$ (c) $\mu = 15, \sigma = 8$

(d) $\mu = 25, \sigma = 2$ (e) $\mu = 25, \sigma = 5$ (f) $\mu = 25, \sigma = 8$

(g) $\mu = 35, \sigma = 2$ (h) $\mu = 35, \sigma = 5$ (i) $\mu = 35, \sigma = 8$

Figure B.3 – Simple Staircase Simulation with different observer models

C Quest+ Simulation



(a) $\mu = 25, \sigma = 1$ (b) $\mu = 25, \sigma = 2$ (c) $\mu = 25, \sigma = 3$

(d) $\mu = 25, \sigma = 4$ (e) $\mu = 25, \sigma = 5$ (f) $\mu = 25, \sigma = 6$

(g) $\mu = 25, \sigma = 7$ (h) $\mu = 25, \sigma = 8$ (i) $\mu = 25, \sigma = 9$

Figure C.1 – Quest+ Simulation with different observer models

D JND search methods accuracy benchmark

Table D.1 – Benchmark of JND search methods accuracy and efficiency

Method	Setting		Trial numbers	Observer models	σ=2	σ=5	σ=8
	Reversal	Step size					
Simple Staircase	6	1	43.10	μ=15	0.82±0.60	2.97±1.80	6.10±3.00
			33.06	μ=25	0.84±0.65	2.94±1.79	6.22±3.14
			23.27	μ=35	0.84±0.64	2.88±1.84	5.74±2.77
		2	26.69	μ=15	0.83±0.63	2.21±1.67	4.14±2.87
			21.82	μ=25	0.78±0.61	2.13±1.63	4.06±2.78
			16.75	μ=35	0.84±0.60	2.22±1.61	4.08±2.68
		3	20.87	μ=15	0.91±0.64	2.04±1.55	3.53±2.60
			17.55	μ=25	0.82±0.72	1.96±1.53	3.57±2.56
			14.17	μ=35	0.83±0.69	1.97±1.59	3.39±2.52
	7	1	45.28	μ=15	0.68±0.53	2.61±1.68	5.68±2.89
			35.24	μ=25	0.68±0.52	2.53±1.68	5.67±2.82
			25.43	μ=35	0.72±0.57	2.54±1.65	5.01±2.56
		2	28.53	μ=15	0.68±0.52	1.87±1.41	3.68±2.60
			23.56	μ=25	0.72±0.55	1.82±1.39	3.54±2.54
			18.57	μ=35	0.69±0.51	1.82±1.40	3.50±2.39
		3	22.56	μ=15	0.80±0.58	1.75±1.37	3.02±2.22
			19.18	μ=25	0.74±0.63	1.70±1.33	2.99±2.22
			15.88	μ=35	0.78±0.55	1.66±1.27	3.05±2.22
	8	1	46.85	μ=15	0.71±0.53	2.50±1.65	5.38±2.69
			36.96	μ=25	0.67±0.53	2.38±1.56	5.34±2.70
			26.99	μ=35	0.69±0.51	2.50±1.60	4.84±2.42
		2	30.14	μ=15	0.70±0.54	1.83±1.38	3.50±2.49
			25.22	μ=25	0.71±0.51	1.79±1.39	3.53±2.49
			20.19	μ=35	0.68±0.53	1.87±1.35	3.39±2.32
		3	24.12	μ=15	0.79±0.56	1.68±1.25	2.95±2.26
			20.78	μ=25	0.74±0.58	1.72±1.33	2.84±2.12
			17.47	μ=35	0.73±0.55	1.68±1.27	2.90±2.14
Relaxed Binary Search			10.53	μ=15	0.79±0.73	1.82±1.44	2.71±2.18
			10.52	μ=25	0.81±0.74	1.76±1.47	2.66±2.13
			10.52	μ=35	0.84±0.75	1.87±1.57	2.75±2.22
Quest+			10	μ=15	2.23±1.69	5.49±3.74	7.36±4.50
				μ=25	1.59±1.02	3.46±2.36	4.75±3.14
				μ=35	1.13±1.14	3.00±2.11	3.92±2.64
			15	μ=15	1.47±1.09	4.16±2.95	6.80±4.23
				μ=25	1.12±0.83	2.93±2.06	4.79±3.10
				μ=35	0.96±0.76	2.72±1.87	4.02±2.60
			20	μ=15	1.15±0.78	3.30±2.32	6.14±4.04
				μ=25	0.96±0.66	2.69±1.72	4.31±2.66
				μ=35	0.81±0.61	2.47±1.60	3.89±2.44
			25	μ=15	0.98±0.64	2.79±1.92	5.46±3.95
				μ=25	0.88±0.61	2.37±1.41	3.95±2.58
				μ=35	0.73±0.54	2.45±1.44	3.63±2.29
			30	μ=15	0.92±0.58	2.47±1.69	4.92±3.44
				μ=25	0.83±0.55	2.30±1.39	3.67±2.34
				μ=35	0.70±0.50	2.35±1.40	3.41±2.08
			35	μ=15	0.88±0.52	2.30±1.57	4.75±3.32
				μ=25	0.83±0.54	2.15±1.18	3.64±2.22
				μ=35	0.65±0.49	2.24±1.29	3.42±1.96
			40	μ=15	0.87±0.51	2.21±1.37	4.50±3.03
				μ=25	0.83±0.50	2.12±1.18	3.44±1.98
				μ=35	0.62±0.44	2.23±1.19	3.33±1.81

E Intra campaign observer behavior analysis

(a) ID: 7 (b) ID: 9 (c) ID: 10 (d) ID: 11

(e) ID: 12 (f) ID: 13 (g) ID: 14 (h) ID: 19

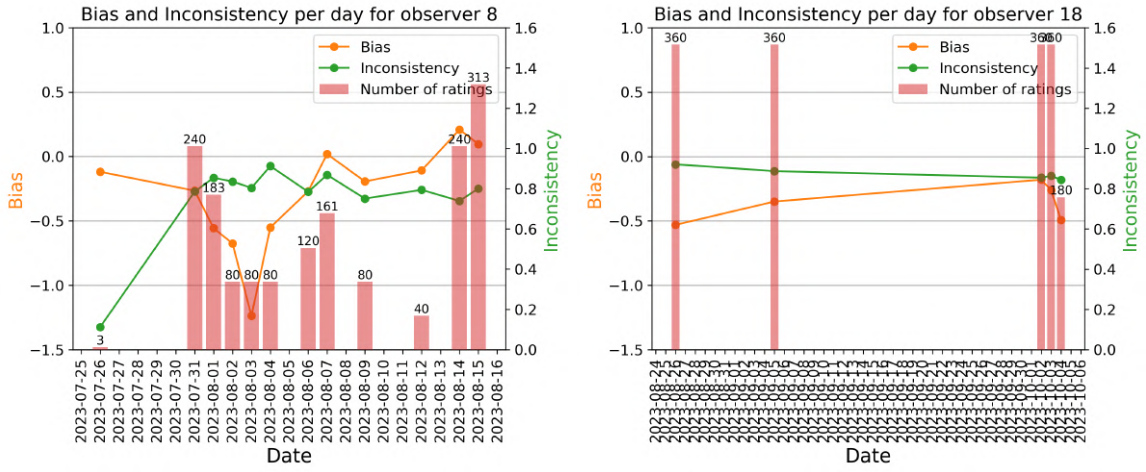
(i) ID: 20 (j) ID: 21 (k) ID: 22 (l) ID: 27

(m) ID: 28 (n) ID: 29 (o) ID: 30 (p) ID: 31

(q) ID: 32 (r) ID: 33 (s) ID: 34 (t) ID: 36

Figure E.1 – Intra campaign observer behavior analysis for naive observers

Besides the 20 observers in Figure E.1, there are also two experts which conducted the same campaign. Because they are experts, the limitation per day was set higer than the naive observers. Their results are as follows:



(a) ID: 8 (b) ID: 18

Figure E.2 – Intra campaign observer behavior analysis for expert

F Details of the Confidence Interval of the MLE

In this appendix, we explain how to compute the confidence interval of the Maximum Likelihood Estimation (MLE), taking the Normal distribution as an example. This method is applicable to other distributions as well. The VW-JND of each video content clip m can be regarded as a random variable J^m. The annotations from the JND subjective test by N observers can be viewed as N independent and identically distributed (*i.i.d.*) samples of J^m. From the observed values $\mathbf{j}^m = [j_1^m, j_2^m, ..., j_N^m]$, we can estimate the parameters of the distribution of J^m using MLE.

We assumed that J^m follows a Gaussian distribution [138, 137, 135, 152, 151, 63]. The probability density function of the Gaussian distribution for video content clip m is given by:

$$f^m(j|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{j-\mu}{\sigma}\right)^2}, \tag{7.3}$$

where μ and σ are two parameters of Gaussian distribution. The likelihood function of clip m:

$$\begin{aligned} L^m(\mu,\sigma^2|\mathbf{j}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\sum_{i=1}^{N}\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2} \end{aligned} \tag{7.4}$$

where j_i^m is the VW-JND value of the i-th observer obtained from the subjective test for clip m. The log-likelihood function of clip m:

$$\begin{aligned} \ell^m(\mu,\sigma^2) &= \log(L^m(\mu,\sigma^2|\mathbf{j})) \\ &= \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\sum_{i=1}^{N}\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2}\right) \\ &= N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\sum_{i=1}^{N}\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2}\right) \\ &= N\log(1) - N\log\left(\sqrt{2\pi\sigma^2}\right) + \log\left(e^{-\sum_{i=1}^{N}\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2}\right) \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\left(\sigma^2\right) - \sum_{i=1}^{N}\frac{1}{2}\left(\frac{j_i^m-\mu}{\sigma}\right)^2 \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(j_i^m-\mu)^2 \end{aligned} \tag{7.5}$$

The estimated parameters of the Gaussian distribution for clip m are the values of μ and σ^2 that maximize the log-likelihood function.

The gradient vector of the log-likelihood function of clip m:

$$\mathbf{u}(\theta) = \frac{\partial \ell^m(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \mu}\ell^m(\mu, \sigma^2) \\ \frac{\partial}{\partial \sigma^2}\ell^m(\mu, \sigma^2) \end{pmatrix}, \tag{7.6}$$

where $\mathbf{u}(\theta) \in \mathbb{R}^{p\times 1}$, p is the number of the parameters. For Gaussian distribution, $p = 2$. For Gaussian distribution:

$$\begin{aligned} \frac{\partial}{\partial \mu}\ell^m(\mu, \sigma^2) &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(j_i^m - \mu\right)^2\right) \\ &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(\left(j_i^m\right)^2 - 2\mu j_i^m + \mu^2\right)\right) \\ &= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}\left(-2j_i^m\right) + 2N\mu\right) \\ &= -\frac{1}{2\sigma^2}\left(-2\sum_{i=1}^{N} j_i^m + 2N\mu\right) \end{aligned} \tag{7.7}$$

$$\frac{\partial}{\partial \sigma^2}\ell^m(\mu, \sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}\left(j_i^m - \mu\right)^2 \tag{7.8}$$

The optimal $\hat{\mu}$ and $\hat{\sigma}^2$ maximize the log-Likelihood function, *i.e.,* $\hat{\mu}$ and $\hat{\sigma}^2$ that make the score vector equal to zero:

$$\mathbf{u}(\theta) = \mathbf{0} \tag{7.9}$$

More specifically, the MLE of the variance of the Gaussian distribution for clip m is:

$$\begin{aligned} \frac{\partial}{\partial \sigma^2}\ell^m(\mu, \sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}\left(j_i^m - \mu\right)^2 = 0 \\ \frac{N}{2\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}\left(j_i^m - \mu\right)^2 \\ \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(j_i^m - \mu\right)^2 \end{aligned} \tag{7.10}$$

The MLE of the mean of the Gaussian distribution for clip m is:

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell^m(\mu,\sigma^2) &= -\frac{1}{2\sigma^2}\left(-2\sum_{i=1}^{N} j_i^m + 2N\mu\right) = 0 \\ 2N\mu &= -2\sum_{i=1}^{N} j_i^m \\ \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} j_i^m\end{aligned} \tag{7.11}$$

The estimation of the μ and σ^2 are based on the samples of the VW-JND values $J^m = [j_1^m, j_2^m, ..., j_N^m]$ from the subjective test. Due to the limited number of samples, the MLE of the parameters may not be accurate. The confidence interval of the MLE can be computed to provide the range of the parameters. The confidence interval of the MLE can be computed using the Fisher's information [22].

The Fisher's information [38] $\boldsymbol{\mathcal{I}}(\theta)$ is the negative expectation of the second derivative of the log likelihood function:

$$\boldsymbol{\mathcal{I}}(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ell^m(\theta)\right) \tag{7.12}$$

where θ is the parameters to be estimated, and $\ell^m(\theta)$ is the log-likelihood function of clip m. For Gaussian distribution, the Fisher's information is a $p \times p$ matrix, where p is the number of the parameters, *i.e.*, $p = 2$ for Gaussian distribution. The Fisher's information matrix for Gaussian distribution is:

$$\boldsymbol{\mathcal{I}}(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ell^m(\theta)\right) = -E\begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\ell^m(\mu,\sigma^2) & \frac{\partial^2}{\partial\mu\partial\sigma^2}\ell^m(\mu,\sigma^2) \\ \frac{\partial^2}{\partial(\sigma^2)\partial\mu}\ell^m(\mu,\sigma^2) & \frac{\partial^2}{\partial(\sigma^2)^2}\ell^m(\mu,\sigma^2) \end{pmatrix} \tag{7.13}$$

The $p \times p$ matrix $\boldsymbol{I}(\theta)$ is called *observed* Fisher's information matrix which is different from the *expected* Fisher's information matrix without the expectation operator:

$$\boldsymbol{I}(\theta) = -\left(\frac{\partial^2}{\partial\theta^2}\ell^m(\theta)\right) = -\begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\ell^m(\mu,\sigma^2) & \frac{\partial^2}{\partial\mu\partial\sigma^2}\ell^m(\mu,\sigma^2) \\ \frac{\partial^2}{\partial(\sigma^2)\partial\mu}\ell^m(\mu,\sigma^2) & \frac{\partial^2}{\partial(\sigma^2)^2}\ell^m(\mu,\sigma^2) \end{pmatrix} \tag{7.14}$$

where each element of the *observed* Fisher's information matrix is:

$$\frac{\partial^2}{\partial\mu^2}\ell^m(\mu,\sigma^2) = -\frac{N}{\sigma^2} \tag{7.15}$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ell^m(\mu,\sigma^2) = \frac{N}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{N}(j_i^m-\mu)^2 \tag{7.16}$$

$$\begin{aligned}\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell^m(\mu,\sigma^2) &= \frac{\partial}{\partial\mu}\left(\frac{\partial}{\partial(\sigma^2)}\ell^m(\mu,\sigma^2)\right)\\ &= \frac{1}{2(\sigma^2)^2}\times\frac{\partial}{\partial\mu}\sum_{i=1}^{N}(j_i^m-\mu)^2\\ &= \frac{1}{2(\sigma^2)^2}\times\left(-2\sum_{i=1}^{N}(j_i^m-\mu)\right)\\ &= -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{N}(j_i^m-\mu)\end{aligned} \tag{7.17}$$

$$\begin{aligned}\frac{\partial^2}{\partial(\sigma^2)\,\partial\mu}\ell^m(\mu,\sigma^2) &= \frac{\partial}{\partial(\sigma^2)}\left(\frac{\partial}{\partial\mu}\ell^m(\mu,\sigma^2)\right)\\ &= \frac{\partial}{\partial(\sigma^2)}\left(-\frac{1}{2\sigma^2}\left(-2\sum_{i=1}^{N}j_i^m+2N\mu\right)\right)\\ &= \frac{1}{(\sigma^2)^2}\left(-\sum_{i=1}^{N}j_i^m+N\mu\right)\\ &= -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{N}(j_i^m-\mu)\end{aligned} \tag{7.18}$$

It can be observed that the cross partial derivative with respect to μ and σ^2 is the same no matter the order of the derivative, which is consistant with the Schwarz theorem [45]. The *observed* Fisher's information matrix for Gaussian distribution is:

$$\boldsymbol{I}(\theta) = \begin{pmatrix} \frac{N}{\sigma^2} & \frac{1}{(\sigma^2)^2}\sum_{i=1}^{N}(j_i^m-\mu) \\ \frac{1}{(\sigma^2)^2}\sum_{i=1}^{N}(j_i^m-\mu) & \frac{1}{(\sigma^2)^3}\sum_{i=1}^{N}(j_i^m-\mu)^2 - \frac{N}{2(\sigma^2)^2} \end{pmatrix} \tag{7.19}$$

It can be observed that the *observed* Fisher's information matrix is actually the negative of the Hessian matrix of the log-likelihood function.

The expected Fisher's information matrix is the expectation of the *observed* Fisher's information matrix:

$$\boldsymbol{\mathcal{I}}(\theta) = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N\sigma^2}{(\sigma^2)^3} - \frac{N}{2(\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2(\sigma^2)^2} \end{pmatrix} \tag{7.20}$$

The inverse of the expected Fisher's information matrix $\boldsymbol{\mathcal{I}}(\theta)^{-1}$:

$$\boldsymbol{\mathcal{I}}(\theta)^{-1} = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\left(\sigma^2\right)^2}{N} \end{pmatrix} \tag{7.21}$$

Similar to the Central Limit Theorem (CLT), if N is large and the J^m are i.i.d. random variables, the distribution of estimated parameters by MLE is asymptotically normal, with the mean equal to the true parameter value, and the variance-covariance matrix equal to the inverse of the expected Fisher's information matrix, denoted as $\boldsymbol{\mathcal{I}}(\theta)^{-1}$. In this matrix, the diagonal elements represent the variances of the estimated parameters, while the off-diagonal elements represent the covariances of the estimated parameters. Because we can not know the true parameters, we use the MLE estimated parameters to replace the true parameters to compute the confidence interval. The 95% confidence interval of the parameters can be computed as:

$$\hat{\theta} \pm 1.96\sqrt{Diag(\boldsymbol{\mathcal{I}}(\hat{\theta}))^{-1}} \tag{7.22}$$

where $\hat{\theta}$ is the estimated parameters by MLE, and $Diag(\boldsymbol{\mathcal{I}}(\hat{\theta}))^{-1}$ is the diagonal elements of the inverse of the expected Fisher's information matrix. More spacificaly, the 95% confidence interval of the mean μ and the variance σ^2 of the Gaussian distribution are:

$$\hat{\mu} \pm 1.96\sqrt{\frac{\hat{\sigma}^2}{N}} \tag{7.23}$$

$$\hat{\sigma} \pm 1.96\sqrt{\frac{2\hat{\sigma}^2}{N}} \tag{7.24}$$

It can be observed that the confidence interval of the MLE is related to the sample size N. The larger the sample size, the smaller the confidence interval. The confidence interval of the MLE can be used to evaluate the accuracy of the MLE estimation. Besides, the confidence interval is related to the variance of the distribution. The larger the variance, the larger the confidence interval. The confidence interval of the MLE can be used to evaluate the reliability of the MLE estimation. This method is applicable to other distributions as well.

G Videos with higest and lowest ΔVMAF$_{\mathbf{SUR}(75\%)}$

In this annex, we showcase the frames of videos with the highest and lowest ΔVMAF$_{\mathrm{SUR}(75\%)}$ values in VideoSet.

G.1 Videos with the 3 highest ΔVMAF$_{\mathbf{SUR}(75\%)}$

Following are the I frames of SRC#120 , SRC#77, SRC#107.

Figure G.1 – I frames of SRC#120, ΔVMAF$_{\mathrm{SUR}(75\%)}$=100 - 75.2246 = 24.7754

Figure G.2 – I frames of SRC#77, ΔVMAF$_{\mathrm{SUR}(75\%)}$=100 - 76.3779 = 23.6221

Figure G.3 – I frames of SRC#107, ΔVMAF$_{\mathrm{SUR}(75\%)}$=100 - 78.4399 = 21.5601

Interestingly, SRC#120 and SRC#77 are both scenes with lawns, while SRC#107 is a scene with turbulent water. They all contain a lot of details and textures. It's notable that for these videos, in the compressed versions where VMAF gives very low scores (up to 75), human eyes cannot easily perceive distortion.

G.2 Videos with the 3 lowest $\Delta\text{VMAF}_{\text{SUR}(75\%)}$

following are the I frames of SRC#15(99.9676), SRC#65(99.9582), SRC#175(99.9558)

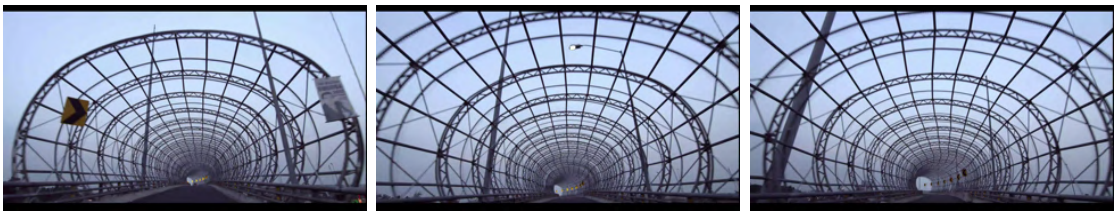


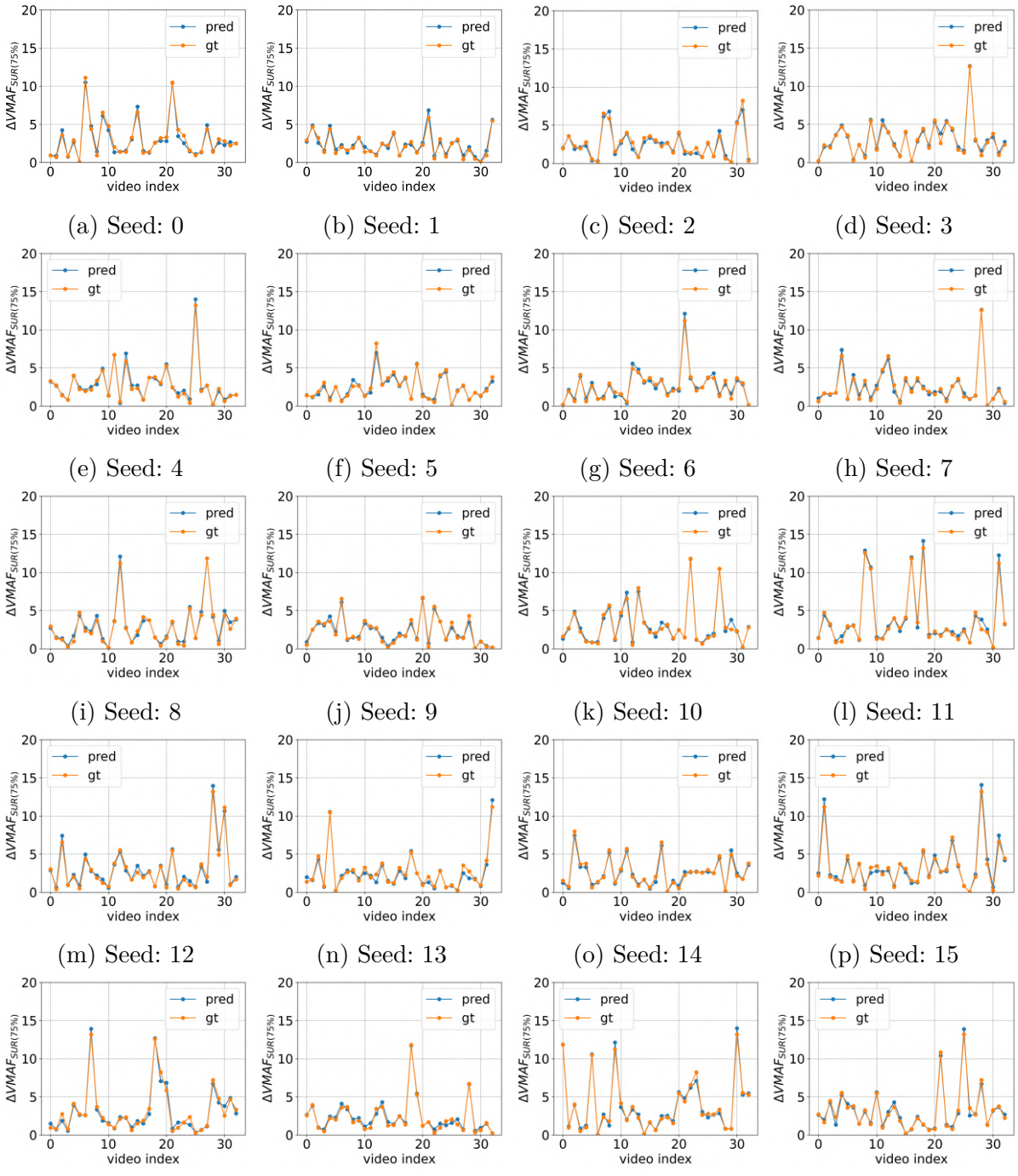
Figure G.4 – I frames of SRC#15, $\Delta\text{VMAF}_{\text{SUR}(75\%)}$=100 - 99.9676 = 0.0324

Figure G.5 – I frames of SRC#65, $\Delta\text{VMAF}_{\text{SUR}(75\%)}$=100 - 99.9582 = 0.0418

Figure G.6 – I frames of SRC#175, $\Delta\text{VMAF}_{\text{SUR}(75\%)}$=100 - 99.9558 = 0.0442

For these videos, in the compressed versions where VMAF gives very high scores (close to 100), human eyes can easily perceive distortion compared to the pristine videos. The common point of these videos is that they have camera movement in the scene, resulting in a lot of motion.

H $\Delta\text{VMAF}_{\text{SUR}(75\%)}$ prediction results with different seeds



(a) Seed: 0 (b) Seed: 1 (c) Seed: 2 (d) Seed: 3

(e) Seed: 4 (f) Seed: 5 (g) Seed: 6 (h) Seed: 7

(i) Seed: 8 (j) Seed: 9 (k) Seed: 10 (l) Seed: 11

(m) Seed: 12 (n) Seed: 13 (o) Seed: 14 (p) Seed: 15

(q) Seed: 16 (r) Seed: 17 (s) Seed: 18 (t) Seed: 19

Figure H.1 – $\Delta\text{VMAF}_{\text{SUR(75\%)}}$ prediction results with different seeds on AMZ-HDR-VJND datasets. (Refer to Chapter 5.3)

I Spatial and Temporal Randomness

Spatial Randomness measures the regularity/randomness of video content in spatial domain. Specifically, videos with low regularity/high randomness spatially will mask the perception of distortion by video compression. SR is measured based on the spatial prediction error [50]. For one frame of video patch, the prediction error of each pixel is calculated as follows:

$$E(i,j) = |y(i,j) - C_{YX}C_X\mathbf{x}(i,j)|\,, \tag{7.25}$$

where $y(i,j)$ is the ground truth value of the pixel (i,j), $\mathbf{x}$ is the neighboring pixels, C_{YX} and C_X are the local properties of the image block, which is a small patch of the entire frame as shown in Figure I.1.

$$\mathbf{X_b} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x}_N]^T \tag{7.26}$$

$$\mathbf{Y_b} = [y_1, y_2, ..., y_N] \tag{7.27}$$

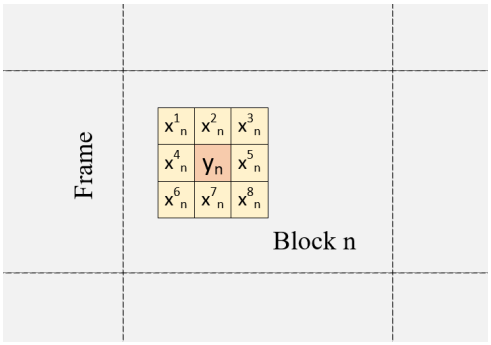


Figure I.1 – Illustration of local properties calculation of image block for Spatial Randomness (SR) computation

$\mathbf{X_b}$ and $\mathbf{Y_b}$ are all the pixels and its neighboring pixels groups (orange and yellow squares in Figure 3.2). The local properties of the current block are calculated using the correlation matrix of $\mathbf{X_b}$ and the cross-correlation matrix of $\mathbf{X_b}$ and $\mathbf{Y_b}$:

$$C_{YX} = \frac{1}{N-1} Y_b X_b^T, \tag{7.28}$$

$$C_X = \frac{1}{N-1} X_b X_b^T. \tag{7.29}$$

The product of local properties and the neighboring pixels vector $\mathbf{x}$ is thus the prediction

of $y(i,j)$. The average of the prediction error $E(i,j)$ of all frames and then the average of each frame for entire videos is the Spatial Randomness for video.

$$SR = \underset{F \in V}{Mean} \left(\underset{i,j \in F}{Mean} \left(E(i,j) \right) \right) \tag{7.30}$$

In our experiment, the block size is set to be the same size as the video patch (W = 360, H = 180)[114]. SR is calculated on the three RGB channels respectively and then mean value of them is calculated. The visualization of Spatial Randomness is shown in Figure I.2. It can be seen that for a video content with low spatial complexity (*e.g.,* SRC037), the SR value is relatively low compared to a video content with high spatial complexity (*e.g.,* SRC089). We can observed that the SR is high in the edges of object in images, it is because the edges is more difficult for linear prediction from neighborhood pixel and the prediction error is relatively high compared with smooth area.

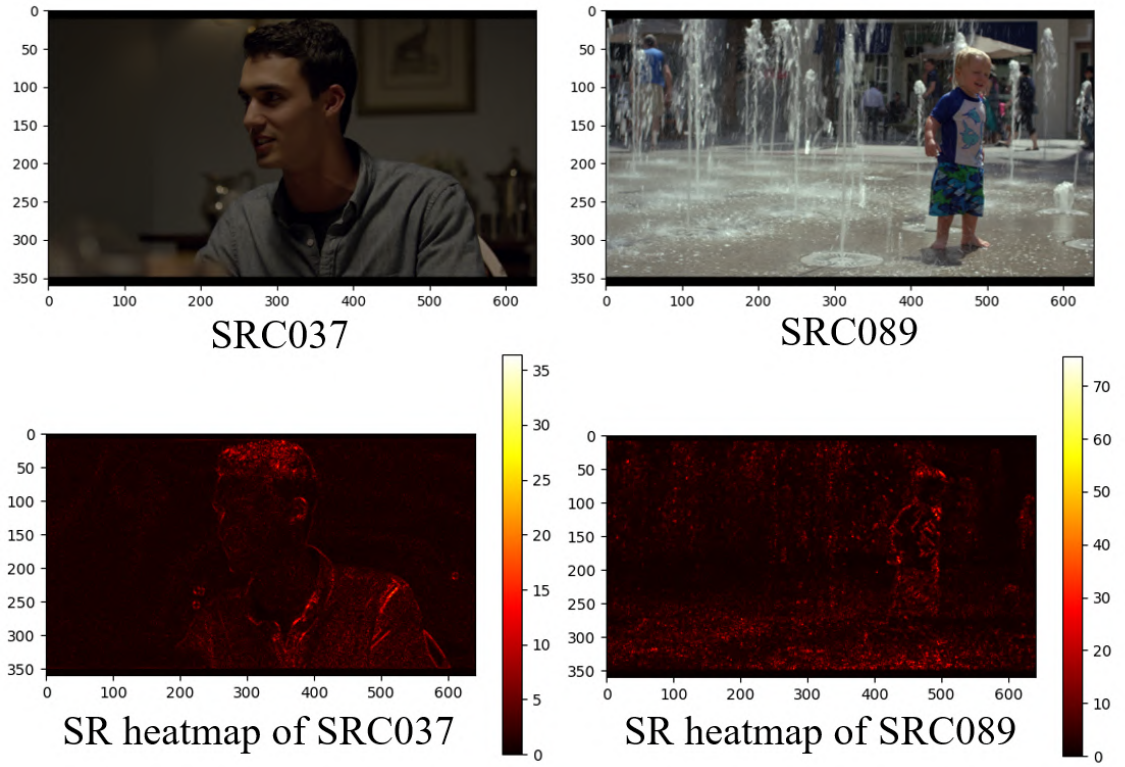


Figure I.2 – Visualization of Spatial Randomness (SR) for two different video content

Temporal Randomness measures the regularity/randomness of video content in temporal domain. Similarly with spatial randomness, video with low regularity/high randomness will mask the perception of distortion by the human eye. The prediction error of one frame from its previous frames is estimated to measure Temporal Randomness

(TR). $Y_m^n \in \mathbb{R}^{l\times(n-m+1)}$ denotes a video sequence from frame n to frame m, each element $y \in \mathbb{R}^{l\times 1}$ is a column vector of a frame with flattened pixel value ($l = H \times W$).

$$Y_m^n = [y(m), y(m+1), ..., y(n)] \tag{7.31}$$

The prediction error is calculated on the basis of a dynamic system [51] :

$$E(n+1) = |y(n+1) - C(n)A(n)x(n)|\,, \tag{7.32}$$

where the $y(n+1)$ is the ground truth value of the frame $n+1$ and $C(n)$, $A(n)$, $x(n)$ can be estimated as follows:

$$Y_m^n = U\Sigma V^T \in \mathbb{R}^{l\times(n-m+1)} \tag{7.33}$$

$$C(n) = U \in \mathbb{R}^{l\times k} \tag{7.34}$$

$$X_m^n = \Sigma V^T \in \mathbb{R}^{k\times(n-m-1)} \tag{7.35}$$

$$A(n) = X_m^n (X_{m-1}^{n-1})^{-1} \in \mathbb{R}^{k\times k} \tag{7.36}$$

Eq. (7.33) is the singular value decomposition of Y_m^n, in which the singular values of the diagonal matrix are sorted from largest to smallest. It should be noticed that in Eq. (7.36), the inverse of matrix X_{m-1}^{n-1} is the Moore-Penrose pseudo-inverse [129] of a matrix X_{m-1}^{n-1}. $x(n) \in \mathbb{R}^{k\times 1}$ is the last column of X_m^n.

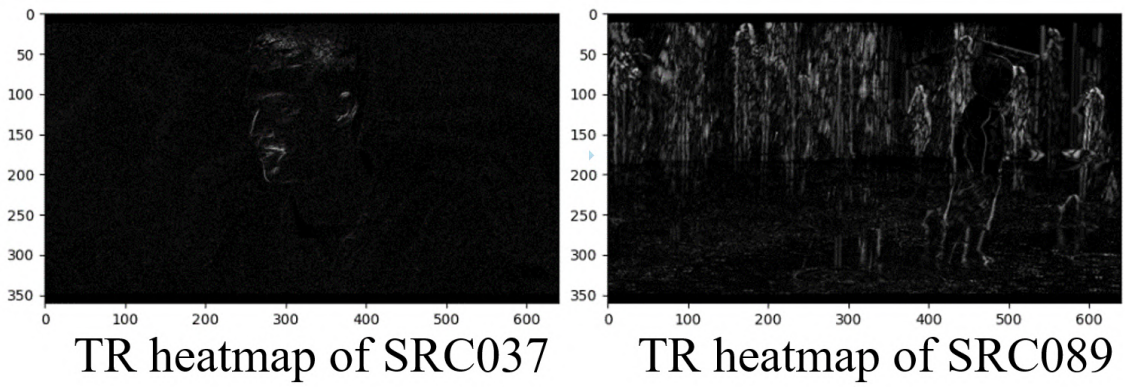


Figure I.3 – Visualization of Temporal Randomness (TR) for two different video content

In our experiment, $(n-m+1) = 5$, which means 5 previous frames are used to predict the next frame. The three RGB channels are used for the TR computation. Visualization of TR is shown in Figure I.3. The same examples as TR are used here. It can be seen that

the video with high motion/less regular content (*e.g.,*) SRC089 has a higher TR than the video content which is more regular (*e.g,* SRC037). TR for one video is the average of prediction error for all frames (except for the first previous p frames, in our cases $p = 5$).

J Contributions to the Scientific Community

— **Organizing Committee**: Served as the Student Volunteer Chair for ACM IMX 2023.

— **Reviewer**: Provided peer reviews for multiple prestigious conferences, including ICIP 2022, ICIP 2023, ICIP 2024, ICASSP 2023, ICME 2023, ICME 2024, and IMX 2023.

— **Teaching**: Taught an Image Processing course for second-year undergraduate students at Nantes Université in 2022 and 2023.

— **Mentorship**: Mentored a master's student for a six-month research project at Nantes Université in 2022.

BIBLIOGRAPHY

[1] Ali Ak et al., « On Spammer Detection In Crowdsourcing Pairwise Comparison Tasks: Case Study On Two Multimedia Qoe Assessment Scenarios », *in*: July 2021, pp. 1–6, DOI: 10.1109/ICMEW53276.2021.9455992 (cit. on p. 56).

[2] Ali Ak et al., « RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images », *in*: *IEEE Transactions on Multimedia* (2022), pp. 1–12, DOI: 10.1109/TMM.2022.3203211 (cit. on pp. 3, 56).

[3] Ali Ak et al., « Video Consumption in Context: Influence of Data Plan Consumption on QoE », *in*: *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 2023, pp. 320–324 (cit. on p. 19).

[4] Critina Amati, Niloy J. Mitra, and Tim Weyrich, « A Study of Image Colourfulness », *in*: *Workshop on Computational Aesthetics*, Aug. 2014, pp. 23–31 (cit. on p. 53).

[5] Hadi Amirpour et al., « VCD: Video Complexity Dataset », *in*: *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, ISBN: 9781450392839, DOI: 10.1145/3524273.3532892 (cit. on pp. 118, 133).

[6] Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, « DeepStream: Video Streaming Enhancements using Compressed Deep Neural Networks », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* (2022), pp. 1–1, ISSN: 1051-8215, 1558-2205, DOI: 10.1109/TCSVT.2022.3229079, (visited on 12/20/2022) (cit. on pp. 92, 118).

[7] Hadi Amirpour, Raimund Schatz, and Christian Timmerer, « Between Two and Six? Towards Correct Estimation of JND Step Sizes for VMAF-based Bitrate Laddering », *in*: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2022, pp. 1–4 (cit. on pp. 86, 94–98).

[8] Hadi Amirpour, Christian Timmerer, and Mohammad Ghanbari, « PSTR: Per-Title Encoding Using Spatio-Temporal Resolutions », *in*: *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2021, pp. 1–6 (cit. on pp. 34, 92, 117, 118).

[9] Hadi Amirpour et al., « A Real-Time Video Quality Metric for HTTP Adaptive Streaming », *in*: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 3810–3814 (cit. on p. 84).

[10] Hadi Amirpour et al., « Exploring Bitrate Costs for Enhanced User Satisfaction: A Just Noticeable Difference (JND) Perspective », *in*: *2024 Data Compression Conference (DCC)*, IEEE, 2024, pp. 432–441 (cit. on p. 117).

[11] Anastasia Antsiferova et al., « Video compression dataset and benchmark of learning-based video-quality metrics », *in*: *Advances in Neural Information Processing Systems*, ed. by S. Koyejo et al., vol. 35, Curran Associates, Inc., 2022, pp. 13814–13825, URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/59ac9f01ea2f701310f3d42037546e4a-Paper-Datasets_and_Benchmarks.pdf (cit. on p. 93).

[12] Apple, *HTTP Live Streaming (HLS) Authoring Specification for Apple Devices | Apple Developer Documentation*, 2015, URL: https://developer.apple.com/documentation/http_live_streaming/http_live_streaming_hls_authoring_specification_for_apple_devices (visited on 06/06/2022) (cit. on pp. 92, 118).

[13] Nabajeet Barman, Nabeel Khan, and Maria G Martini, « Analysis of spatial and temporal information variation for 10-bit and 8-bit video sequences », *in*: *2019 IEEE 24th international workshop on computer aided modeling and design of communication links and networks (CAMAD)*, IEEE, 2019, pp. 1–6 (cit. on p. 15).

[14] Peter GJ Barten, *Contrast sensitivity of the human eye and its effects on image quality*, SPIE press, 1999 (cit. on p. 28).

[15] G Békésy, « VON (1947). A new audiometer », *in*: *Acta Oto-laryngologica* () (cit. on pp. 4, 43).

[16] Alexandre Benoit et al., « Quality assessment of stereoscopic images », *in*: *EURASIP journal on image and video processing* 2008 (2009), pp. 1–13 (cit. on p. 33).

[17] Abdelhak Bentaleb et al., « A survey on bitrate adaptation schemes for streaming media over HTTP », *in*: *IEEE Communications Surveys & Tutorials* 21.*1* (2018), pp. 562–585 (cit. on pp. 34, 92).

[18] Dimitri Bertsekas and John N Tsitsiklis, *Introduction to probability*, vol. 1, Athena Scientific, 2008 (cit. on p. 68).

[19] Philip R Bevington et al., « Data reduction and error analysis for the physical sciences », *in*: *Computers in Physics* 7.*4* (1993), pp. 415–416 (cit. on p. 62).

[20] Madhukar Bhat, Jean-Marc Thiesse, and Patrick Le Callet, « Combining Video Quality Metrics To Select Perceptually Accurate Resolution In A Wide Quality Range: A Case Study », *in*: *IEEE ICIP*, Sept. 2021, pp. 2164–2168, ISBN: 978-1-66544-115-5, DOI: 10.1109/ICIP42928.2021.9506310, URL: https://ieeexplore.ieee.org/document/9506310/ (visited on 04/27/2023) (cit. on p. 119).

[21] G. Bjontegaard, « Calculation of average PSNR differences between RD-curves », *in*: *VCEG-M33* (2001) (cit. on p. 134).

[22] Ralph Allan Bradley and Milton E Terry, « Rank analysis of incomplete block designs: I. The method of paired comparisons », *in*: *Biometrika* 39.*3/4* (1952), pp. 324–345 (cit. on pp. 73, 156).

[23] Kjell Brunnström and Marcus Barkowsky, « Statistical quality of experience analysis: on planning the sample size and statistical significance testing », *in*: *Journal of Electronic Imaging* 27.*5* (2018), pp. 053013–053013 (cit. on p. 33).

[24] Tianqi Chen and Carlos Guestrin, « Xgboost: A scalable tree boosting system », *in*: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794 (cit. on p. 96).

[25] Manri Cheon et al., « Ambiguity of objective image quality metrics: A new methodology for performance evaluation », *in*: *Signal Processing: Image Communication* 93 (2021), p. 116150 (cit. on pp. 5, 66).

[26] Tom N Cornsweet, « Changes in the appearance of stimuli of very high luminance. », *in*: *Psychological review* 69.*4* (1962), p. 257 (cit. on p. 43).

[27] Scott Daly, « Digital images and human vision », *in*: *ch. The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity* (1993), pp. 179–206 (cit. on p. 28).

[28] Pierre David et al., « Estimating Uncertainty On Video Quality Metrics », *in*: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5 (cit. on pp. 5, 66).

[29] Jan De Cock et al., « Complexity-Based Consistent-Quality Encoding in the Cloud », *in*: *IEEE ICIP*, Sept. 2016, pp. 1484–1488, DOI: 10.1109/ICIP.2016.7532605 (cit. on pp. 92, 118, 134).

[30] Frederic Dufaux et al., *High dynamic range video: from acquisition, to display and applications*, Academic Press, 2016 (cit. on p. 41).

[31] Walter H Ehrenstein and Addie Ehrenstein, « Psychophysical methods », *in*: *Modern techniques in neuroscience research*, Springer, 1999, pp. 1211–1241 (cit. on pp. 42, 46).

[32] GÖSta Ekman, « Weber's law and related functions », *in*: *The Journal of Psychology* 47.*2* (1959), pp. 343–352 (cit. on p. 34).

[33] Chunling Fan et al., « Interactive subjective study on picture-level just noticeable difference of compressed stereoscopic images », *in*: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 8548–8552 (cit. on pp. 37, 38).

[34] Chunling Fan et al., « SUR-Net: Predicting the satisfied user ratio curve for image compression with deep learning », *in*: *2019 eleventh international conference on quality of multimedia experience (QoMEX)*, IEEE, 2019, pp. 1–6 (cit. on pp. 59, 62).

[35] Gustav Theodor Fechner, *Elemente der psychophysik*, vol. 2, Breitkopf u. Härtel, 1860 (cit. on pp. 4, 42).

[36] William Feller, *An introduction to probability theory and its applications, Volume 2*, vol. 81, John Wiley & Sons, 1991 (cit. on p. 68).

[37] Francesc J Ferri et al., « Comparative study of techniques for large-scale feature selection », *in*: *Machine Intelligence and Pattern Recognition*, vol. 16, Elsevier, 1994, pp. 403–413 (cit. on p. 110).

[38] Ronald A Fisher, « On the mathematical foundations of theoretical statistics », *in*: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.*594-604* (1922), pp. 309–368 (cit. on p. 156).

[39] Ronald Aylmer Fisher, *Statistical methods for research workers*, Springer, 1992 (cit. on p. 66).

[40] Franz Götz-Hahn et al., « KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild », *in*: *IEEE Access* 9 (2021), pp. 72139–72160 (cit. on pp. 15, 20).

[41] Gerald J Hahn and William Q Meeker, *Statistical intervals: a guide for practitioners*, vol. 92, John Wiley & Sons, 2011 (cit. on p. 69).

[42] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein, « Textural features for image classification », *in*: *IEEE Transactions on systems, man, and cybernetics 6* (1973), pp. 610–621 (cit. on pp. 53, 109).

[43] N B Harikrishnan et al., « Comparative evaluation of image compression techniques », *in*: *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 2017, pp. 1–4, DOI: 10.1109/ICAMMAET.2017.8186637 (cit. on p. 131).

[44] David Hasler and Sabine E Suesstrunk, « Measuring colorfulness in natural images », *in*: *Human vision and electronic imaging VIII*, vol. 5007, SPIE, 2003, pp. 87–95 (cit. on pp. 53, 94, 109).

[45] JR Hicks, *Mathematical Analysis for Economists*, 1939 (cit. on p. 157).

[46] David M Hoffman and Dale Stolitzka, « A new standard method of subjective assessment of barely visible image artifacts and a new public database », *in*: *Journal of the Society for Information Display* 22.*12* (2014), pp. 631–643 (cit. on p. 40).

[47] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger, « SOS: The MOS is not enough! », *in*: *2011 third international workshop on quality of multimedia experience*, IEEE, 2011, pp. 131–136 (cit. on pp. 5, 20, 21, 66).

[48] Tobias Hoßfeld et al., « Quantification of YouTube QoE via crowdsourcing », *in*: *2011 IEEE International Symposium on Multimedia*, IEEE, 2011, pp. 494–499 (cit. on p. 3).

[49] Vlad Hosu et al., *The Konstanz Natural Video Database*, 2017, URL: http://database.mmsp-kn.de (cit. on p. 33).

[50] Sudeng Hu et al., « Compressed image quality metric based on perceptually weighted distortion », *in*: *IEEE Transactions on Image Processing* 24.*12* (2015), pp. 5594–5608 (cit. on pp. 2, 100, 104, 108, 162).

[51] Sudeng Hu et al., « Objective video quality assessment based on perceptually weighted mean squared error », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 27.*9* (2016), pp. 1844–1855 (cit. on pp. 100, 104, 108, 164).

[52] Qin Huang et al., « Measure and prediction of HEVC perceptually lossy/lossless boundary QP values », *in*: *2017 data compression conference (DCC)*, IEEE, 2017, pp. 42–51 (cit. on pp. 38, 39).

[53] Quan Huynh-Thu and Mohammed Ghanbari, « Scope of validity of PSNR in image/video quality assessment », *in*: *Electronics letters* 44.*13* (2008), pp. 800–801 (cit. on p. 33).

[54] Quan Huynh-Thu et al., « Study of rating scales for subjective quality assessment of high-definition video », *in*: *IEEE Transactions on Broadcasting* 57.*1* (2010), pp. 1–14 (cit. on p. 33).

[55] ITU, *Subjective video quality assessment methods for multimedia applications*, 2008 (cit. on pp. 16, 17, 53).

[56] ITU-R, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Recommendation BT.500-14, 2019 (cit. on pp. 1, 18, 19, 22, 33, 56, 57).

[57] ITU-R, *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, ITU-R Recommendation Recommendation P.913, 2021 (cit. on pp. 18, 19, 56–58).

[58] ITU-R, *Subjective video quality assessment methods for multimedia applications*, ITU-R Recommendation Recommendation P.910, 2022 (cit. on pp. 1, 18, 33, 56, 58, 94).

[59] ITU-R BT.1769, « Parameter values for an expanded hierarchy of LSDI image formats for production and international programme exchange », *in*: *Int'l Telecommunication Union* (2006) (cit. on p. 18).

[60] ITU-R BT.2013-1, « A reference viewing environment for evaluation of HDTV program material or completed programmes », *in*: *Int'l Telecommunication Union* (2013) (cit. on p. 18).

[61] P ITU-T RECOMMENDATION, « Subjective video quality assessment methods for multimedia applications », *in*: (1999) (cit. on pp. 3, 109).

[62] Carlos M Jarque and Anil K Bera, « A test for normality of observations and regression residuals », *in*: *International Statistical Review/Revue Internationale de Statistique* (1987), pp. 163–172 (cit. on p. 101).

[63] Mohsen Jenadeleh et al., « Crowdsourced Estimation of Collective Just Noticeable Difference for Compressed Video with the Flicker Test and QUEST+ », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* (2024) (cit. on pp. 2, 15, 37–39, 43, 44, 47, 71, 123, 147, 148, 154).

[64] Lina Jin et al., « Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis », *in*: *Electronic Imaging* 2016.*13* (2016), pp. 1–9 (cit. on pp. 37, 38).

[65] Satu Jumisko-Pyykkö and Miska M. Hannuksela, « Does Context Matter in Quality Evaluation of Mobile Television? », *in*: MobileHCI '08, Amsterdam, The Netherlands: Association for Computing Machinery, 2008, pp. 63–72, ISBN: 9781595939524, DOI: 10.1145/1409240.1409248, URL: https://doi.org/10.1145/1409240.1409248 (cit. on p. 19).

[66] Andreas Kah et al., « Fundamental relationships between subjective quality, user acceptance, and the VMAF metric for a quality-based bit-rate ladder design for over-the-top video streaming services », *in*: *Applications of Digital Image Processing XLIV*, vol. 11842, SPIE, 2021, pp. 316–325 (cit. on pp. 94, 97, 98).

[67] Angeliki V. Katsenou et al., « VMAF-Based Bitrate Ladder Estimation for Adaptive Streaming », *in*: *PCS*, June 2021, pp. 1–5, DOI: 10.1109/PCS50896.2021.9477469 (cit. on pp. 93, 118).

[68] Sehwan Ki et al., « Learning-based just-noticeable-quantization-distortion modeling for perceptual video coding », *in*: *IEEE Transactions on Image Processing* 27.*7* (2018), pp. 3178–3193 (cit. on p. 32).

[69] F Kozamernik et al., « SAMVIQ—A new EBU methodology for video quality evaluations in multimedia », *in*: *SMPTE motion imaging journal* 114.*4* (2005), pp. 152–160 (cit. on p. 33).

[70] Lukáš Krasula et al., « On the accuracy of objective image and video quality models: New methodology for performance evaluation », *in*: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2016, pp. 1–6 (cit. on p. 66).

[71] Lukáš Krasula et al., « Subjective video quality for 4K HDR-WCG content using a browser-based approach for" at-home" testing », *in*: *Electronic Imaging* 35 (2023), pp. 263–1 (cit. on pp. 14, 20).

[72] Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al., « Qualinet white paper on definitions of quality of experience », *in*: *European network on quality of experience in multimedia systems and services (COST Action IC 1003)* 3.*2012* (2012) (cit. on p. 18).

[73] Junghyuk Lee et al., « A perception-based framework for wide color gamut content selection », *in*: *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 709–713 (cit. on pp. 53, 112).

[74] Junghyuk Lee et al., « Wide color gamut image content characterization: method, evaluation, and applications », *in*: *IEEE Transactions on Multimedia* 23 (2020), pp. 3817–3827 (cit. on p. 53).

[75] Jing Li et al., « A Probabilistic Graphical Model for Analyzing the Subjective Visual Quality Assessment Data from Crowdsourcing », *in*: *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 3339–3347, ISBN: 9781450379885, DOI: 10.1145/3394171.3413619, URL: https://doi.org/10.1145/3394171.3413619 (cit. on p. 56).

[76] Jing Li et al., *GPM: A Generic Probabilistic Model to Recover Annotator's Behavior and Ground Truth Labeling*, 2020, DOI: 10.48550/ARXIV.2003.00475, URL: https://arxiv.org/abs/2003.00475 (cit. on p. 56).

[77] Jing Li et al., « Quantifying the Influence of Devices on Quality of Experience for Video Streaming », *in*: *2018 Picture Coding Symposium (PCS)*, 2018, pp. 308–312, DOI: 10.1109/PCS.2018.8456304 (cit. on pp. 18, 19, 23, 33).

[78] Zhi Li and Christos G Bampis, « Recover subjective quality scores from noisy measurements », *in*: *2017 Data compression conference (DCC)*, IEEE, 2017, pp. 52–61 (cit. on p. 58).

[79] Zhi Li et al., « A Simple Model for Subject Behavior in Subjective Experiments », *in*: *Human Vision and Electronic Imaging 2020, Burlingame, CA, USA, 26-30 January 2020*, Ingenta, 2020, DOI: 10.2352/ISSN.2470-1173.2020.11.HVEI-

131, URL: https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-131 (cit. on pp. 53, 56–58).

[80] Zhi Li et al., « Toward a practical perceptual video quality metric », *in*: *The Netflix Tech Blog* 6.*2* (2016) (cit. on pp. 34, 35, 84, 85, 100, 118).

[81] Zhuoran Li et al., « AVC, HEVC, VP9, AVS2 or AV1? — A Comparative Study of State-of-the-Art Video Encoders on 4K Videos », en, *in*: *Image Analysis and Recognition*, vol. 11662, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 162–173, ISBN: 978-3-030-27201-2 978-3-030-27202-9, DOI: 10.1007/978-3-030-27202-9_14, URL: http://link.springer.com/10.1007/978-3-030-27202-9_14 (visited on 10/26/2023) (cit. on p. 123).

[82] Liang Liao et al., « Exploring the effectiveness of video perceptual representation in blind video quality assessment », *in*: *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 837–846 (cit. on p. 20).

[83] Harris R Lieberman and Alex P Pentland, « Microcomputer-based estimation of psychophysical thresholds: the best PEST », *in*: *Behavior Research Methods & Instrumentation* 14.*1* (1982), pp. 21–25 (cit. on p. 43).

[84] Hanhe Lin et al., « Large-scale crowdsourced subjective assessment of picturewise just noticeable difference », *in*: *IEEE transactions on circuits and systems for video technology* 32.*9* (2022), pp. 5859–5873 (cit. on pp. 38, 39, 42).

[85] Hanhe Lin et al., « Subjective assessment of global picture-wise just noticeable difference », *in*: *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2020, pp. 1–6 (cit. on p. 32).

[86] Hanhe Lin et al., « SUR-FeatNet: Predicting the satisfied user ratio curve for image compression with deep feature learning », *in*: *Quality and User Experience* 5 (2020), pp. 1–23 (cit. on pp. 59, 62).

[87] Joe Yuchieh Lin et al., « Experimental design and analysis of JND test on coded image/video », *in*: *Applications of digital image processing XXXVIII*, vol. 9599, SPIE, 2015, pp. 324–334 (cit. on pp. 34, 37).

[88] Suiyi Ling et al., « Towards better quality assessment of high-quality videos », *in*: *Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*, 2020, pp. 3–9 (cit. on pp. 16, 53, 109).

[89] Suiyi Ling et al., « Towards Perceptually-Optimized Compression Of User Generated Content (UGC): Prediction Of UGC Rate-Distortion Category », *in*: *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6, DOI: 10.1109/ICME46284.2020.9102752 (cit. on p. 53).

[90] Huanhua Liu et al., « Deep learning-based picture-wise just noticeable distortion prediction model for image compression », *in*: *IEEE Transactions on Image Processing* 29 (2019), pp. 641–656 (cit. on p. 32).

[91] Jiawen Liu, Jingwen Zhu, and Patrick Le Callet, « Bridge the Gap between Visual Difference Prediction Model and Just Noticeable Difference Subjective Datasets », *in*: *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2023, pp. 1–5 (cit. on p. 86).

[92] Xiaohua Liu et al., « JND-Pano: Database for just noticeable difference of JPEG compressed panoramic images », *in*: *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, Springer, 2018, pp. 458–468 (cit. on pp. 37, 38).

[93] Yaxuan Liu et al., « The First Comprehensive Dataset with Multiple Distortion Types for Visual Just-Noticeable Differences », *in*: *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 2820–2824 (cit. on p. 40).

[94] Rafał Mantiuk et al., « HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions », *in*: *ACM Transactions on graphics (TOG)* 30.*4* (2011), pp. 1–14 (cit. on p. 28).

[95] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk, « Comparison of four subjective methods for image quality assessment », *in*: *Computer graphics forum*, vol. 31, 8, Wiley Online Library, 2012, pp. 2478–2491 (cit. on p. 34).

[96] Rafał K. Mantiuk et al., « FovVideoVDP: a visible difference predictor for wide field-of-view video », en, *in*: *ACM Transactions on Graphics* 40.*4* (Aug. 2021), pp. 1–19, ISSN: 0730-0301, 1557-7368, DOI: 10.1145/3450626.3459831, URL: https://dl.acm.org/doi/10.1145/3450626.3459831 (visited on 06/11/2023) (cit. on p. 85).

[97] Vignesh V Menon et al., « INCEPT: Intra CU Depth Prediction for HEVC », *in*: *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6, DOI: 10.1109/MMSP53017.2021.9733517 (cit. on p. 131).

[98] Vignesh V Menon et al., « JND-aware Two-pass Per-title Encoding Scheme for Adaptive Live Streaming », *in*: (2023), DOI: 10.36227/techrxiv.22256704.v1, URL: https://www.techrxiv.org/articles/preprint/JND-aware_Two-pass_Per-title_Encoding_Scheme_for_Adaptive_Live_Streaming/22256704 (cit. on p. 131).

[99] Vignesh V Menon et al., *Video Quality Assessment with Texture Information Fusion for Streaming Applications*, 2023, arXiv: 2302.14465 [cs.MM], URL: https://arxiv.org/abs/2302.14465 (cit. on p. 131).

[100] Vignesh V Menon et al., « Energy-Efficient Multi-Codec Bitrate-Ladder Estimation for Adaptive Video Streaming », *in*: *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2023, pp. 1–5 (cit. on p. 84).

[101] Vignesh V Menon et al., « Just noticeable difference-aware per-scene bitrate-laddering for adaptive video streaming », *in*: *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 1673–1678 (cit. on p. 117).

[102] Vignesh V Menon et al., « Optimal quality and efficiency in adaptive live streaming with JND-aware low latency encoding », *in*: *Proceedings of the 3rd Mile-High Video Conference*, 2024, pp. 61–67 (cit. on p. 84).

[103] Vignesh V Menon et al., « Perceptually-Aware Per-Title Encoding for Adaptive Video Streaming », *in*: *IEEE ICME*, July 2022, pp. 1–6, DOI: 10.1109/ICME52920.2022.9859744 (cit. on pp. 93, 119, 120, 131).

[104] Vignesh V Menon et al., « VCA: video complexity analyzer », *in*: *Proceedings of the 13th ACM multimedia systems conference*, 2022, pp. 259–264 (cit. on pp. 94, 131, 133).

[105] Vignesh V. Menon, « Video Coding Enhancements for HTTP Adaptive Streaming », *in*: *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6905–6909, ISBN: 9781450392037, DOI: 10.1145/3503161.3548753 (cit. on p. 131).

[106] Babak Naderi and Ross Cutler, « A crowdsourcing approach to video quality assessment », *in*: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 2810–2814 (cit. on p. 3).

[107] E Nakasu et al., « A statistical analysis of MPEG-2 picture quality for television broadcasting », *in*: *SMPTE journal* 105.*11* (1996), pp. 702–711 (cit. on p. 1).

[108] Sanaz Nami et al., *Lightweight Multitask Learning for Robust JND Prediction using Latent Space and Reconstructed Frames*, 2024 (cit. on p. 36).

[109] Sanaz Nami et al., « MTJND: Multi-task deep learning framework for improved JND prediction », *in*: *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 1245–1249 (cit. on pp. 36, 101).

[110] Nandhu Nandhakumar, « Recent activities of the ultra hd forum », *in*: *SMPTE Motion Imaging Journal* 131.*8* (2022), pp. 107–110 (cit. on p. 41).

[111] Manish Narwaria et al., « An objective method for high dynamic range source content selection », *in*: *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2014, pp. 13–18 (cit. on p. 53).

[112] Yana Nehmé et al., « Comparison of subjective methods for quality assessment of 3D graphics in virtual reality », *in*: *ACM Transactions on Applied Perception (TAP)* 18.*1* (2020), pp. 1–23 (cit. on p. 33).

[113] *Netflix Public Dataset.https://github.com/Netflix/vmaf/ blob/master/resource/doc/datasets.md*, Online; accessed 20-Feb-2023 (cit. on p. 59).

[114] Alexandre Ninassi et al., « Considering temporal variations of spatial visual distortions in video quality assessment », *in*: *IEEE Journal of Selected Topics in Signal Processing* 3.*2* (2009), pp. 253–265 (cit. on p. 163).

[115] J. Ozer, « Finding the Just Noticeable Difference with Netflix VMAF », *in*: `https://streaminglearningcenter.com/codecs/finding-the-just-noticeable-difference-with-netflix-vmaf.html` (cit. on pp. 86, 93, 97, 98).

[116] Adrien Paire, Anne Hillairet de Boisferon, and Céline Paeye, « Empirical validation of QUEST+ in PSE and JND estimations in visual discrimination tasks », *in*: *Behavior Research Methods* 55.*8* (2023), pp. 3984–4001 (cit. on p. 47).

[117] Maria Perez-Ortiz et al., « From pairwise comparisons and rating to a unified quality scale », *in*: *IEEE Transactions on Image Processing* 29 (2019), pp. 1139–1151 (cit. on p. 40).

[118] Anne-Flore Perrin et al., « When is the Cleaning of Subjective Data Relevant to Train UGC Video Quality Metrics? », *in*: *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1466–1470, DOI: 10.1109/ICIP46576.2022.9897997 (cit. on pp. 57, 65).

[119] Margaret H Pinson, *Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality*, tech. rep., Institute for Telecommunication Sciences, 2020 (cit. on pp. 5, 66).

[120] Margaret H. Pinson, Marcus Barkowsky, and Patrick Le Callet, « Selecting scenes for 2D and 3D subjective video quality tests », *in*: *EURASIP Journal on Image and Video Processing* 2013.*1* (Aug. 2013), pp. 50–61 (cit. on p. 16).

[121] Alexander Raake et al., « Scalable video quality model for ITU-T P. 1203 (aka P. NATS) for bitstream-based monitoring of HTTP adaptive streaming », *in*: *Proc. QoMEX*, 2017 (cit. on p. 109).

[122] Rakesh Rao Ramachandra Rao et al., « Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation », *in*: *QoMEX*, Athlone, Ireland, May 2020 (cit. on p. 109).

[123] Yuriy A. Reznik et al., « Optimal Design of Encoding Profiles for ABR Streaming », *in*: *PV Workshop*, June 2018, pp. 43–47, ISBN: 978-1-4503-5773-9, DOI: 10.1145/3210424.3210436, URL: https://doi.org/10.1145/3210424.3210436 (visited on 03/28/2022) (cit. on pp. 93, 119).

[124] Werner Robitza, *CRF guide (Constant rate factor in x264, x265 and libvpx)*, https://slhck.info/video/2017/02/24/crf-guide.html, Feb. 2017 (cit. on p. 41).

[125] Graeme D Ruxton, « The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test », *in*: *Behavioral Ecology* 17.*4* (2006), pp. 688–690 (cit. on p. 21).

[126] Xuelin Shen et al., « Just noticeable distortion profile inference: A patch-level structural visibility learning approach », *in*: *IEEE Transactions on Image Processing* 30 (2020), pp. 26–38 (cit. on pp. 32, 38, 39).

[127] Zeina Sinno and Alan Conrad Bovik, « Large-scale study of perceptual video quality », *in*: *IEEE Transactions on Image Processing* 28.*2* (2018), pp. 612–627 (cit. on p. 33).

[128] Alex J Smola and Bernhard Schölkopf, « A tutorial on support vector regression », *in*: *Statistics and computing* 14.*3* (2004), pp. 199–222 (cit. on p. 104).

[129] Gilbert Strang, *Linear algebra and its applications*. Belmont, CA: Thomson, Brooks/Cole, 2006 (cit. on p. 164).

[130] Farzad Tashtarian et al., « LALISA: Adaptive Bitrate Ladder Optimization in HTTP-based Adaptive Live Streaming », *in*: *IEEE/IFIP NOMS*, IEEE, May 2023, pp. 1–9, ISBN: 978-1-66547-716-1, DOI: 10.1109/NOMS56928.2023.10154347, URL: https://ieeexplore.ieee.org/document/10154347/ (visited on 07/25/2023) (cit. on pp. 93, 119, 120).

[131] Tao Tian et al., « Just noticeable difference level prediction for perceptual image compression », *in*: *IEEE Transactions on Broadcasting* 66.*3* (2020), pp. 690–700 (cit. on p. 32).

[132] Praveen Kumar Tiwari et al., « Accelerating x265 with Intel® Advanced Vector Extensions 512 », *in*: *White Paper on the Intel Developers Page* (2018), URL: https://www.intel.com/content/dam/develop/external/us/en/documents/mcw-intel-x265-avx512.pdf (cit. on p. 134).

[133] Amos Tversky, « Elimination by Aspects: A Theory of Choice », *in*: *Psychological Review* 79.*4* (1972), pp. 281–299, DOI: 10.1037/h0032955 (cit. on p. 23).

[134] V. V. Menon et al., *Transcoding Quality Prediction for Adaptive Video Streaming*, 2023, arXiv: 2304.10234 [cs.MM], URL: https://arxiv.org/abs/2304.10234 (cit. on p. 131).

[135] Haiqiang Wang et al., « Analysis and prediction of JND-based video quality model », *in*: *2018 Picture Coding Symposium (PCS)*, IEEE, 2018, pp. 278–282 (cit. on pp. 32, 34, 36, 62, 67, 71, 100, 147, 154).

[136] Haiqiang Wang et al., « MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset », *in*: *2016 IEEE international conference on image processing (ICIP)*, IEEE, 2016, pp. 1509–1513 (cit. on pp. 34, 38, 39, 101).

[137] Haiqiang Wang et al., « Prediction of satisfied user ratio for compressed video », *in*: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 6747–6751 (cit. on pp. 6, 32, 34, 36, 59, 67, 71, 100, 103, 108, 154).

[138] Haiqiang Wang et al., « VideoSet: A large-scale compressed video quality dataset based on JND measurement », *in*: *Journal of Visual Communication and Image Representation* 46 (2017), pp. 292–302 (cit. on pp. 2–4, 34, 35, 37–39, 43, 44, 46, 47, 52, 67, 71, 72, 74, 86, 92, 94, 100, 101, 103, 123, 147, 148, 154).

[139] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, « Multiscale structural similarity for image quality assessment », *in*: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, Ieee, 2003, pp. 1398–1402 (cit. on pp. 84, 85).

[140] Zhou Wang et al., « Image quality assessment: from error visibility to structural similarity », *in*: *IEEE transactions on image processing* 13.*4* (2004), pp. 600–612 (cit. on pp. 84, 85).

[141] AB Watson and DG Pelli, « QUEST: A Bayesian adaptive psychophysical method », *in*: *Perception* (1990) (cit. on p. 43).

[142] Andrew B Watson, « QUEST+: A general multidimensional Bayesian adaptive psychometric method », *in*: *Journal of Vision* 17.*3* (2017), pp. 10–10 (cit. on pp. 4, 43).

[143] Florian Wickelmaier and Christian Schmid, « A Matlab function to estimate choice model parameters from paired-comparison data », *in*: *Behavior Research Methods, Instruments, & Computers* 36 (2004), pp. 29–40 (cit. on p. 24).

[144] Adam Wieckowski et al., « VVenC: An Open And Optimized VVC Encoder Implementation », *in*: *IEEE ICMEW*, Shenzhen, China: IEEE, July 2021, pp. 1–2, ISBN: 978-1-66544-989-2, DOI: 10.1109/ICMEW53276.2021.9455944, URL: https://ieeexplore.ieee.org/document/9455944/ (visited on 05/15/2023) (cit. on p. 125).

[145] Stefan Winkler, « On the properties of subjective ratings in video quality experiments », *in*: *2009 International Workshop on Quality of Multimedia Experience*, IEEE, 2009, pp. 139–144 (cit. on p. 33).

[146] Stephen Wolf and Margaret Pinson, « Video quality measurement techniques », *in*: *2002.* (2002) (cit. on p. 109).

[147] Haoning Wu et al., « Exploring video quality assessment on user generated contents from aesthetic and technical perspectives », *in*: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20144–20154 (cit. on p. 20).

[148] Ping-Hao Wu et al., « Encoding parameters prediction for convex hull video encoding », *in*: *2021 picture coding symposium (PCS)*, IEEE, 2021, pp. 1–5 (cit. on p. 93).

[149] Junyong You and Jari Korhonen, « Deep Neural Networks for No-Reference Video Quality Assessment », *in*: *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2349–2353, DOI: 10.1109/ICIP.2019.8803395 (cit. on p. 130).

[150] Richard Zhang et al., « The unreasonable effectiveness of deep features as a perceptual metric », *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595 (cit. on p. 40).

[151] Xinfeng Zhang et al., « Satisfied-user-ratio modeling for compressed video », *in*: *IEEE Transactions on Image Processing* 29 (2020), pp. 3777–3789 (cit. on pp. 32, 34, 59, 62, 71, 100, 147, 154).

[152] Yun Zhang et al., « Deep learning based just noticeable difference and perceptual quality prediction models for compressed video », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 32.*3* (2021), pp. 1197–1212 (cit. on pp. 36, 59, 62, 67, 71, 101, 147, 154).

[153] Jingwen Zhu and Patrick Le Callet, « Just noticeable difference (JND) and satisfied user ratio (SUR) prediction for compressed video: research proposal », *in*: *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, pp. 393–397 (cit. on p. 32).

[154] Jingwen Zhu, Anne-Flore Perrin, and Patrick Le Callet, « Subjective test methodology optimization and prediction framework for Just Noticeable Difference and Satisfied User Ratio for compressed HD video », *in*: *2022 Picture Coding Symposium (PCS)*, IEEE, 2022, pp. 313–317 (cit. on pp. 31, 59, 62, 65, 136).

[155] Jingwen Zhu et al., « A framework to map vmaf with the probability of just noticeable difference between video encoding recipes », *in*: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, IEEE, 2022, pp. 1–5 (cit. on pp. 16, 31, 51, 123).

[156] Jingwen Zhu et al., « Beyond Curves and Thresholds - Introducing Uncertainty Estimation to Satisfied User Ratios for Compressed Video », *in*: *2024 Picture Coding Symposium (PCS)*, IEEE, 2024 (cit. on pp. 55, 83).

[157] Jingwen Zhu et al., « Elevating Your Streaming Experience with Just Noticeable Difference (JND)-based Encoding », *in*: *Proceedings of the 2nd Mile-High Video Conference*, 2023, pp. 128–129 (cit. on p. 119).

[158] Jingwen Zhu et al., « Enhancing Satisfied User Ratio (SUR) Prediction for VMAF Proxy through Video Quality Metrics », *in*: *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2023, pp. 1–5 (cit. on p. 83).

[159] Jingwen Zhu et al., « On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video », *in*: *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 4213–4217 (cit. on pp. 34, 59, 83, 108).

[160] Jingwen Zhu et al., « Subjective Test Environments: A Multifaceted Examination of Their Impact on Test Results », *in*: *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, 2023, pp. 298–302 (cit. on pp. 11, 83).

[161] Jingwen Zhu et al., « ZREC : robust recovery of mean and percentile opinion scores », working paper or preprint, Mar. 2023, URL: `https://hal.science/hal-04017583` (cit. on pp. 20, 22, 55).

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Title: High-end Video Streaming Quality in the Wild: Measuring and Predicting Satisfied User Ratio

Keywords: Just Noticeable Difference, Satisfied User Ratio, Streaming Optimization, Video Quality Assessment

Abstract: The human eye cannot perceive small pixel changes in images or videos until a certain threshold of distortion. In the context of video compression, Just Noticeable Difference (JND) is the smallest distortion level from which the human eye can perceive the difference between reference video and the distorted/compressed one. Satisfied-User-Ratio (SUR) curve is the (complementary) cumulative distribution function of the individual JNDs of a viewer group. JND and SUR have been widely investigated for compressed image and video to use the least resources without damaging the Quality of Experience. In this thesis, we introduce a new AtHome protocol for subjective studies, which combines in-lab and crowdsourcing methodologies. We optimize JND search methods, reducing subjective test time, and collect new JND datasets for HD-SDR and UHD-HDR videos. We improve data reliability with a screening method named ZREC and propose methods for estimating confidence intervals for SUR. We further conduct a longitudinal study based on the AtHome protocol. We develop a pipeline to predict SUR using VQMs as proxy and parameter-driven models to predict SUR using encoding parameters as proxy, enhancing practicality for streaming services. Finally, we demonstrate how integrating JND and SUR into bitrate ladder optimization can save bitrate and storage.

Titre : Qualité du Streaming Vidéo Haut de Gamme dans des Conditions Réelles : Mesurer et Prédire le Taux d'Utilisateurs Satisfaits

Mot clés : Différence Juste Perceptible, Taux d'Utilisateurs Satisfaits, Optimisation du Streaming, Évaluation de la Qualité Video

Résumé : L'œil humain ne peut percevoir de petits changements de pixels dans les images ou les vidéos jusqu'à ce qu'un certain seuil de distorsion soit atteint. Dans le contexte de la compression vidéo, la Différence Juste Perceptible (JND) est le plus petit niveau de distorsion à partir duquel l'œil humain peut percevoir la différence entre une vidéo de référence et la vidéo déformée/compressée. La courbe du Taux d'Utilisateurs Satisfaits (SUR) est la fonction de distribution cumulative (complémentaire) des JND individuels d'un groupe de observateurs. Les JND et SUR ont été largement étudiés pour les images et vidéos compressées afin d'utiliser les ressources minimales sans compromettre la Qualité de l'Expérience. Dans cette thèse, nous introduisons un nouveau protocole AtHome pour les études subjectives, qui combine les approches en laboratoire et de crowdsourcing. Nous optimisons les méthodes de recherche JND, réduisant ainsi le temps des tests subjectifs, et collectons de nouveaux ensembles de données JND pour vidéos HD-SDR et UHD-HDR. Nous améliorons la fiabilité des données avec une méthode appelée ZREC et proposons des méthodes pour estimer les intervalles de confiance pour SUR. Nous menons également une étude longitudinale basée sur le protocole AtHome. Nous développons un pipeline pour prédire SUR en utilisant les VQMs comme proxy et des modèles basés sur les paramètres d'encodage comme proxy, améliorant ainsi la praticité pour les services de streaming. Enfin, nous démontrons comment l'intégration de JND et SUR dans l'optimisation de l'échelle de débit peut économiser le débit et le stockage.