

# Video Quality Assessment for Resolution Cross-Over in Live Sports

Jingwen Zhu<sup>1,2</sup>, Yixu Chen<sup>2</sup>, Hai Wei<sup>2</sup>, Sriram Sethuraman<sup>2</sup>, Yongjun Wu<sup>2</sup>

<sup>1</sup>Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

<sup>2</sup>Amazon Prime Video, Seattle, USA

**Abstract**—In adaptive bitrate streaming, resolution cross-over refers to the point on the convex hull where the encoding resolution should switch to achieve better quality. Accurate cross-over prediction is crucial for streaming providers to optimize resolution at given bandwidths. Most existing works rely on objective Video Quality Metrics (VQM), particularly VMAF, to determine the resolution cross-over. However, these metrics have limitations in accurately predicting resolution cross-overs. Furthermore, widely used VQMs are often trained on subjective datasets collected using the Absolute Category Rating (ACR) methodologies, which we demonstrate introduces significant uncertainty and errors in resolution cross-over predictions. To address these problems, we first investigate different subjective methodologies and demonstrate that Pairwise Comparison (PC) achieves better cross-over accuracy than ACR. We then propose a novel metric, Resolution Cross-over Quality Loss (RCQL), to measure the quality loss caused by resolution cross-over errors. Furthermore, we collected a new subjective dataset (LSCO) focusing on live streaming scenarios and evaluated widely used VQMs, by benchmarking their resolution cross-over accuracy.

**Index Terms**—Adaptive Bitrate Streaming, Video Quality Metric (VQM), Quality of Experience (QoE)

## I. INTRODUCTION

Video streaming has experienced significant growth, driven by the widespread availability of high-speed Internet and the increasing use of mobile devices. It has become an essential part of daily life, serving purposes in entertainment, education, business, and more. To accommodate varying bandwidth and device types among end-users, Adaptive BitRate streaming (ABR) methods [1] are widely used. In ABR, video content is encoded at multiple bitrate-resolution pairs, referred to as representations. These representations form a bitrate ladder [2], allowing video quality to be dynamically adjusted based on the viewer’s available bandwidth and device capabilities.

Fixed bitrate ladders, often used in streaming protocols like HLS, have traditionally been employed. However, this “one-size-fits-all” approach may not be optimal due to the diversity of video content, and it often fails to deliver the best possible quality to end users. To address this, many studies have proposed per-title or per-scene bitrate laddering [3]–[6]. In these approaches,

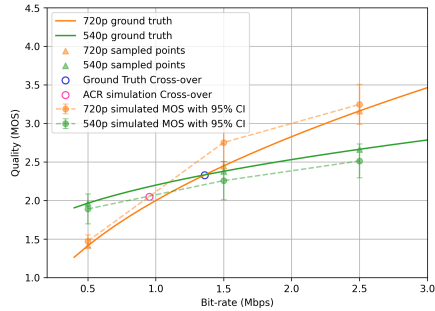


Fig. 1: Illustration of resolution cross-over error introduced by ACR: Solid lines show the ground truth cross-over at 1.4 Mbps between 720p and 540p. Dashed lines represent MOS from ACR simulations, with the cross-over at 900 kbps.

videos are encoded with varying parameters, such as resolution and bitrate, and their quality is evaluated. An optimized bitrate ladder is then constructed by selecting representations from a convex hull derived from the quality measurements of the encoded representations.

Resolution cross-over refers to the point on the convex hull where the encoding resolution should switch to achieve better quality. The solid orange and green lines in Fig. 1 illustrate the resolution cross-over between 720p and 540p. For this video content, if the bitrate is higher than the cross-over point (1.4 Mbps), the representations with 720p encoding resolution will be preferred over 540p. In contrast, 540p representation should be used.

Previous works have attempted to predict the resolution cross-over directly [7] or to use the resolution cross-over to design the bitrate ladder [3]–[6]. These studies primarily rely on Video Quality Metrics (VQMs). While it is well understood that subjective quality assessments are more reliable, they are often too expensive and impractical to perform for every piece of video content, particularly in live streaming scenarios. Consequently, these works use objective VQMs, especially the perceptual quality metric VMAF [8], as a pseudo “ground truth” for determining the resolution cross-over.

However, while VQMs such as VMAF and P1204.3 [9] have shown strong correlations with

subjective video quality, their accuracy in predicting the resolution cross-over compared to subjective assessments remains an open question. Chen *et al.* [10] demonstrated that both VMAF and P1204.3 often fail to predict the resolution cross-over accurately and tend to incorrectly favor higher-resolution videos.

Furthermore, learning-based objective VQMs, such as VMAF, P1204.3, EQM [10] are trained and tested on subjective datasets collected using the Absolute Category Rating (ACR) methodology [8]–[13]. As shown in Fig. 1, our simulation demonstrates that due to the inherent uncertainty in ACR subjective studies, the resolution cross-over derived from ACR studies can significantly deviate from the ground truth assumption (detailed in II-A).

Moreover, to the best of our knowledge, no established metric exists to evaluate the accuracy of resolution cross-over obtained by VQM. For VQM evaluation, the most commonly used metrics are correlations, such as Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC). However, it is evident that a higher correlation with subjective datasets does not necessarily guarantee improved resolution cross-over accuracy.

This paper addresses the challenges associated with resolution cross-over prediction in the context of live streaming by presenting the following contributions:

- 1) We demonstrate errors in resolution cross-over determination using ACR through simulation.
- 2) A pilot study comparing ACR and PC shows that PC provides more accurate resolution cross-over.
- 3) We introduce Resolution Cross-Over Quality Loss (RCQL) to assess cross-over accuracy.
- 4) We collect the Live Sport Cross-Over (LSCO) dataset, propose a method for cleaning incomplete PC data, and benchmark RCQL on LSCO for live streaming use cases.

## II. SUBJECTIVE STUDY DESIGN FOR RESOLUTION CROSS-OVER

### A. ACR and its limitation

ACR is a widely used methodology for video quality evaluation. Participants rate video stimuli on a predefined scale, often with a hidden reference (ACR-HR). A common example is the 5-point scale: 5 (Excellent) to 1 (Bad), with variants like 5-point or 9-point discrete/continuous scales [14]. ACR is popular for being easy to conduct and time-efficient, with results comparable to other methods like DSIS and SAMVIQ [15]. However, ACR has limitations, *e.g.*, subjects may interpret the scale differently, leading to inconsistencies.

We assessed the precision of ACR in determining resolution cross-over points through simulations based

on real data. Ground truth quality values were manually chosen as plausible assumptions. Observer ratings  $R_s$  deviate from these assumptions (ground truth  $\mu_s$ ), following a Gaussian distribution with standard deviation  $\sigma_s$ , influenced by quality range and test environment, as per the SOS hypothesis [16]. Using the HDR-LIVE dataset [17], we modeled  $\sigma_s$  via the SOS curve. Adding  $\sigma_s$  to the ground truth assumptions allowed us to simulate ACR ratings for  $N$  participants ( $N=33$  to be consistent with HDR-LIVE) by sampling from this distribution. The simulated Mean Opinion Score (MOS) was then obtained by averaging these simulated ratings. Fig. 1 shows that due to ACR uncertainty, the cross-over points can significantly deviate from the ground truth.

### B. Pair Comparison with Active Sampling

Considering the errors introduced by ACR for determining the cross-over point, we decided to use alternative test methodologies. Perez-Ortiz *et al.* [18] demonstrated that PC is more accurate method compared to ACR-HR. Similarly, Testolina *et al.* [19] proposed using PC for fine-grained subjective visual quality assessment. We therefore considered PC to collect subjective data for resolution cross-over evaluation.

One major limitation of PC is its  $O(n^2)$  time complexity [20], where  $n$  is the number of candidate stimuli. Several methods [20]–[23] have been proposed to reduce the number of comparisons, among which Active Sampling achieves the best performance [24]. We used the Active Sampling method from [23] for dataset collection. Active Sampling dynamically selects the most informative pairs for comparison instead of evaluating all combinations. The PC results are stored as a Pair Comparison Matrix (PCM). To convert the PCM into a continuous quality scale, we used the `pwcmp` tool [18] to recover the Just Objectable Difference (JOD) scale [25]. This approach relies on Thurstone Case  $V$  assumptions, calibrated so a 1-unit difference on the quality scale corresponds to 75% of observers selecting one video over the other. The JOD reconstruction is performed per video content, with no cross-content comparisons.

To further confirm that PC can provide more accurate cross-over than ACR, we conducted a pilot study using both ACR, PC with active sampling and an expert viewing evaluation. For ACR, we used a 9-point discrete scale [26]. A MOS value of 0 corresponds to “Bad”, and a MOS value of 8 corresponds to “Excellent”. Three 10-second video clips from live streaming applications were selected and encoded at resolutions of 2160p, 1080p, 720p, and 540p, with bitrates ranging from 200 Kbps to 20 Mbps, resulting in 67 encoded videos. The study involved 24 participants for ACR, resulting in a total of  $24 \times 67 = 1,608$  ratings. For PC, the number of active

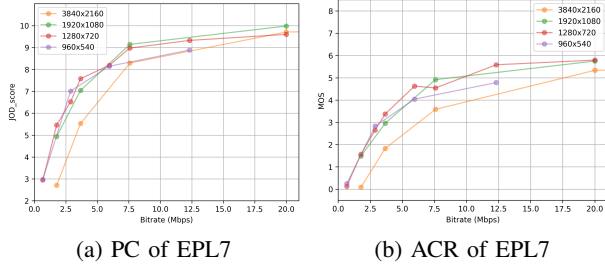


Fig. 2: Pilot study example: Comparison of PC vs ACR

sampling iterations was set to 25 [23], producing 1,650 ratings (30 participants  $\times$  55 ratings each), which is comparable to the number of ratings in ACR. Fig. 2 showcases the MOS from ACR and the JOD from PC obtained in the pilot study for video content EPL7. (Continuous curves are generated using ‘pchip’ interpolation.) It can be observed that the cross-over points derived from PC and ACR differ significantly. Five experts participated in the expert viewing, where they were asked to indicate their preferred resolution for each bitrate of each content in the dataset. The results are presented in Table I. An “x” indicates that no cross-over was observed within the specified bitrate range (200 Kbps to 20 Mbps). For cases with multiple cross-over points, such as the cross-over between 1080p and 720p in EPL7 with ACR (Fig. 2b), we computed the average by taking the mean of the minimum and maximum cross-over bitrates. The cross-over bitrates obtained using PC align more closely with expert evaluations than those from ACR. PC data also shows better correlation with bitrate within the same resolution, following the principle that lower bitrates should not result in better quality within same resolution for the same encoding settings. Table II presents SROCC results, indicating PC achieves better correlation in most cases.

TABLE I: PC and ACR Pilot Study Results Compared with Expert Viewing. The values are in Mbps.

		4k vs 1080p	1080p vs 720p	720p vs 540p
EPL7	Expert	x	5.95	4.67
	PC	x	<b>5.66</b>	<b>2.85</b>
	ACR	x	8.47	2.23
UCL8	Expert	11.87	4.29	1.39
	PC	<b>12.67</b>	<b>3.80</b>	0.84
	ACR	19.46	2.37	<b>1.92</b>
tennis 5	Expert	7.19	1.09	1.39
	PC	<b>6.59</b>	<b>1.47</b>	x
	ACR	14.00	1.77	<b>0.82</b>

### C. LSCO Dataset collection

Based on the results from the pilot study, we conducted a full study for resolution cross-over evaluation using PC with active sampling. The collected dataset is named Live Sport Cross-Over (LSCO). We selected

TABLE II: SROCC between bitrate and quality scores for ACR and PC in the pilot study.

		2160p	1080p	720p	540p
EPL7	PC	1.0000	1.0000	<b>1.0000</b>	1.0000
	ACR	1.0000	1.0000	0.9762	1.0000
UCL8	PC	0.9000	<b>1.0000</b>	<b>0.9762</b>	1.0000
	ACR	0.9000	0.9429	0.9701	1.0000
tennis 5	PC	<b>1.0000</b>	0.9643	<b>1.0000</b>	0.8000
	ACR	0.9000	0.9643	0.9910	<b>1.0000</b>

20 video clips covering a wide range of Spatial Information (SI) and Temporal Information (TI) [26], each approximately 10 seconds long, focusing on live sports events such as football, soccer, and tennis. The selected content was encoded using our in-house encoder into different resolutions, similar to the setup in the pilot study. To minimize participant fatigue, each participant evaluated only 55 pairs, and each session lasted less than 30 minutes. A total of 131 participants contributed to the study, resulting in over 7,000 collected ratings. We used two calibrated OLED55C3PUA 4K displays, with the viewing distance controlled at 1.5 times the screen height following the ITU recommendation [26].

### III. OBSERVER SCREENING

Observer screening is crucial for reliable QoE datasets. While well-established methods exist for ACR and DCR [26], [27], screening for PC data is less developed. Methods like Cohen’s Kappa [28] and RT distance [29], and Ak *et al.*’s ambiguity-weighted RT distance [30], are limited to complete PC data. We propose a new approach to compute inter-observer consistency for incomplete PCMs collected via active sampling. This metric helps identify and filter outliers by considering: 1) pair ambiguity, 2) observer agreement, and 3) the number of ratings per pair.

#### A. Ambiguity

For each observer  $o_i$ , assume that he/she voted on  $N$  pairs of videos, with each pair denoted as  $p_n$ . The number of ratings (i.e., how many people voted on this pair) for  $p_n$  is represented as  $r_n$ . For pair  $p_n$ , the number of ratings for  $A$  is denoted as  $a_n$ , the number of ratings for  $B$  is denoted as  $b_n$ , and the number of ties is denoted as  $t_n$ . We know that:

$$a_n + b_n + t_n = r_n \quad (1)$$

The ambiguity of pair  $p_n$  can be calculate as follow:

$$Ambiguity_{p_n} = \frac{|a_n - b_n|}{r_n} = \frac{|a_n - b_n|}{a_n + b_n + t_n} \quad (2)$$

It is straightforward to observe that the ambiguity of  $p_n$  lies between 0 and 1: **Ambiguity = 0**: This pair is highly ambiguous, with equal votes for  $A$  and  $B$ , or all participants voting for a tie. **Ambiguity = 1**: This pair is not ambiguous at all, with all participants voting exclusively for  $A$  or  $B$ .

### B. Agreement

The agreement of observer  $o_i$  with all other observers for  $p_n$  can be computed with weighted agreement:

$$W_i(p_n) = \begin{cases} P(A)_n = a_n/r_n & \text{if } v_i = A \\ P(B)_n = b_n/r_n & \text{if } v_i = B \\ P(T)_n = t_n/r_n & \text{if } v_i = T \end{cases} \quad (3)$$

where  $v_i$  is the vote of observer  $o_i$  for video pair  $p_n$ .

### C. Inter-observer Consistency

For a given observer  $o_i$ , the basic idea to compute his/her consistency is to calculate the average of the agreement between him/her and other observers across all the pairs he/she voted on. The higher this agreement, the more consistent this observer is with others. However, computing consistency solely based on agreement is not entirely fair, particularly for videos with high ambiguity. To address this, we weight the agreement based on ambiguity, assigning larger weights to less ambiguous pairs and smaller weights to more ambiguous pairs. In active sampling, pairs are rated by varying numbers of observers, so the number of ratings per pair must be considered. Pairs with more votes should be weighted higher, while those rated only once should be excluded from consistency calculations. The inter-observer consistency  $C_i$  can be calculated as:

$$C_i = \frac{\sum_{n=1}^N (r_n - 1) \times \frac{|a_n - b_n|}{r_n} \times W_i(p_n)}{\sum_{n=1}^N (r_n - 1)} \quad (4)$$

The larger  $C_i$  is, the more consistent this observer is with others. To further validate the proposed method, we injected synthetic spammers, and detailed results can be found in Sec. V-A.

## IV. RESOLUTION CROSS-OVER QUALITY LOSS

Evaluating VQM performance often involves measuring their correlation (e.g., PLCC, SROCC) with subjective scores (e.g., MOS, JOD) on study datasets. However, higher correlation does not guarantee better resolution cross-over accuracy, as correlation reflects global agreement but overlooks specific errors at the cross-over point. Fig. 3 illustrates how the RCQL is computed. Suppose there are two resolutions,  $R1$  and  $R2$ , and we aim to determine their resolution cross-over point. The subjective quality is measured, and a function is fitted to describe the corresponding RD curves, denoted as  $f_{R1}(x)$  and  $f_{R2}(x)$  for the two resolutions, respectively. We assume that both  $f_{R1}(x)$  and  $f_{R2}(x)$  are convex and monotonically increasing. Similarly, the predicted video quality metric functions are denoted as  $\hat{f}_{R1}(x)$  and  $\hat{f}_{R2}(x)$ . The bitrate at the cross-over point obtained from the subjective study can be computed as:

$$C_{R1R2} = \min\{x | f_{R1}(x) = f_{R2}(x)\} \quad (5)$$

Similarly, the bitrate at the cross-over point predicted by the VQM can be computed as:

$$\hat{C}_{R1R2} = \min\{x | \hat{f}_{R1}(x) = \hat{f}_{R2}(x)\} \quad (6)$$

The difference between the bitrates of the two cross-over points can be calculated as:

$$\Delta Bitrate = |C_{R1R2} - \hat{C}_{R1R2}| \quad (7)$$

It can be observed in Fig. 3 that in the bitrate range between  $\hat{C}_{R1R2}$  and  $C_{R1R2}$ , the subjective study indicates that  $R2$  provides better quality than  $R1$  and should be selected on the convex hull. However, due to errors in cross-over prediction by the VQM,  $R1$  is incorrectly selected instead of  $R2$  within this bitrate range. The resulting quality loss for users can be quantified as the integral of the subjective quality difference between  $R1$  and  $R2$  over this bitrate range. Mathematically, the RCQL is computed as:

$$RCQL = \left| \int_{C_{R1R2}}^{\hat{C}_{R1R2}} f_{R1}(x) dx - \int_{C_{R1R2}}^{\hat{C}_{R1R2}} f_{R2}(x) dx \right| \quad (8)$$

A higher RCQL value indicates lower cross-over accuracy for the given metric.

Fig. 4 demonstrates why  $\Delta Bitrate$  alone is not sufficient to measure cross-over accuracy. While the  $\Delta Bitrate$  in Case 1 is smaller than in Case 2, the subjective quality difference between  $R1$  and  $R2$  is much smaller in Case 2 compared to Case 1. As a result, the actual quality loss in Case 2 is lower than in Case 1.

Our proposed RCQL metric effectively measures the quality loss caused by selecting the incorrect resolution. It provides a more accurate and reliable measure to reflect the QoE. We also computed the average quality loss over the range of the mistake as:

$$RCQL_{avg} = RCQL_s / \Delta Bitrate \quad (9)$$

This reflects the average quality loss (measured in JOD units) over the mistaken bitrate range.

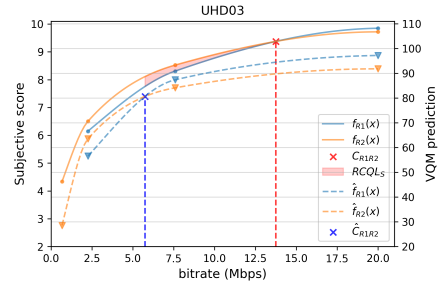


Fig. 3: Illustration of RCQL computation for a VQM between two resolutions,  $R1$  and  $R2$ .

TABLE III: Benchmark RCQL on LSCO datasets

		PSNRy	SSIM	MS-SSIM	VMAF	VMAF 4k	VMAF 4k neg	P1204.3	EQM NR	EQM FR
$\Delta$ Bitrate (Kbps) $\downarrow$	2160p vs 1080p	3175.97	4533.55	3512.11	2930.42	3158.60	3142.59	2472.82	<b>1691.42</b>	2042.62
	1080p vs 720p	1028.79	5852.18	2370.15	<b>845.64</b>	847.35	977.67	1299.44	1692.27	1886.10
	720p vs 540p	<b>308.04</b>	2255.98	671.32	340.51	367.33	350.80	523.37	453.48	267.61
$RCQL_s$ (JOD $\times$ Kbps) $\downarrow$	2160p vs 1080p	1861.03	1440.17	1395.24	1635.23	1792.11	1787.83	1234.36	<b>487.68</b>	635.71
	1080p vs 720p	<b>169.67</b>	3157.68	850.07	199.28	188.66	184.61	299.53	679.59	1189.30
	720p vs 540p	<b>32.12</b>	1474.12	202.78	39.46	53.24	45.96	119.12	120.18	45.34
$RCQL_{avg}$ (JOD) $\downarrow$	2160p vs 1080p	0.2237	0.1716	0.1746	0.1921	0.2051	0.2047	0.1850	<b>0.1091</b>	0.1634
	1080p vs 720p	0.1482	0.5510	0.2619	0.1822	0.1711	0.1685	<b>0.1423</b>	0.2894	0.3289
	720p vs 540p	0.0893	0.3750	0.1707	<b>0.0770</b>	0.0842	0.0815	0.1781	0.1818	0.1186

## V. EXPERIMENTAL RESULTS

### A. Inter-observer consistency

We generated 10 synthetic spammers who provided random ratings and computed the inter-observer consistency together with all participants using the method proposed in Sec. III. The results are presented in Fig. 5. It can be observed that the synthetic spammers (after the green line) exhibit significantly lower inter-observer consistency. This observation also helps establish a threshold to identify outlier observers. In this study, we chose 0.3 as our threshold and removed 2 outliers before proceeding with further analysis.

### B. Benchmark of VQM on our collected LSCO datasets

We compared various VQMs, including PSNR, SSIM [31], MS-SSIM [32], VMAF [8], P1204.3 [9], and EQM [10]. The EQM include a No Reference (NR) model and Full Reference (FR) model. The correlation results are presented in Table IV. It can be observed that EQM outperforms other metrics on the entire dataset as well as across different resolutions. We also evaluated various VQMs on cross-over accuracy using the RCQL metric proposed in Section IV. The results in Table III show that different VQMs perform variably across resolutions. For the cross-over between 2160p and 1080p, EQM NF outperforms other metrics. Interestingly, for lower resolution cross-overs, PSNR achieves better performance.

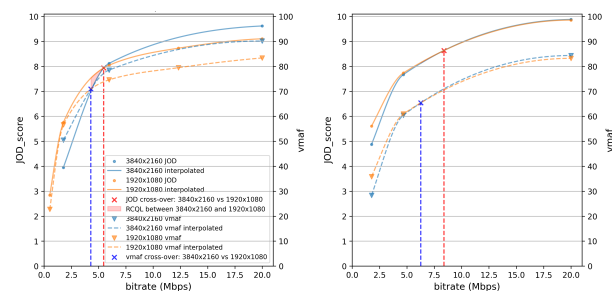


Fig. 4: Demonstration of RCQL and  $\Delta$ Bitrate between VMAF predictions and subjective scores for the resolution cross-over between 2160p and 1080p.

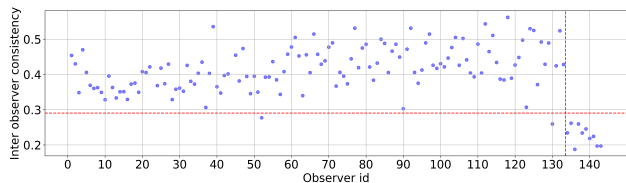


Fig. 5: Inter-observer consistency for each participant, including 10 synthetic spammers (id from 133 to 143)

TABLE IV: Benchmark of correlation between VQM and our collected LSCO datasets

		PSNRy	SSIM	MS-SSIM	VMAF	P1204.3	EQM NR	EQM FR
SROCC	2160p	0.8322	0.9091	0.8322	0.8601	0.9371	<b>0.972</b>	<b>0.972</b>
	1080p	0.7147	0.8765	0.8206	0.8471	0.9176	<b>0.9824</b>	0.9765
	720p	0.7059	0.7941	0.7118	0.8	0.8647	0.9382	<b>0.9588</b>
	540p	0.522	0.5714	0.4341	0.6264	0.6978	<b>0.8571</b>	0.8462
	overall	0.8149	0.8011	0.813	0.8839	0.9321	<b>0.9463</b>	0.9457
PLCC	2160p	0.7611	0.8371	0.7797	0.8334	0.8765	0.9422	<b>0.9504</b>
	1080p	0.7381	0.846	0.7758	0.8408	0.9378	0.9776	<b>0.9801</b>
	720p	0.7356	0.7357	0.7231	0.8392	0.8949	<b>0.9781</b>	0.9778
	540p	0.5506	0.6173	0.5494	0.6912	0.7827	<b>0.9145</b>	0.9074
overall	0.7831	0.7146	0.7359	0.8558	0.9076	0.9525	<b>0.954</b>	

This observation is interesting because it is commonly known that PSNR has a relatively low correlation with subjective study datasets compared to learning-based VQMs such as VMAF and EQM (as also shown in Table IV). However, for resolution cross-over accuracy, PSNR is outperforming both VMAF and EQM on low resolution. This might be because EQM is trained on datasets that focus more on higher resolutions (HD and 4K). Another possible reason is that the resolution cross-over accuracy does not evaluate the cross-content generalization ability of a VQM, which is penalized when using correlation metrics such as SROCC across the entire dataset. This reveals that a higher correlation does not necessarily guarantee better cross-over accuracy.

## VI. CONCLUSION

In this paper, we demonstrated the limitations of ACR in predicting resolution cross-overs and showed that PC can improve cross-over accuracy. We proposed a mathematical measure of inter-observer consistency to identify and remove spammers during subjective studies. We further introduced the RCQL metric to specifically evaluate cross-over accuracy and benchmarked state-of-the-art VQMs on the LSCO dataset under live streaming

scenarios. Experimental results show that higher correlation with subjective scores does not necessarily translate to better cross-over accuracy. Additionally, for resolution cross-overs, EQM performs better at higher resolutions, while PSNR is more effective at lower resolutions. These findings provide valuable insights for selecting appropriate quality metrics when optimizing bitrate ladders in adaptive streaming applications.

## REFERENCES

- [1] Abdelhak Bentaleb, Bayan Taani, Ali C. Begen, Christian Timmerer, and Roger Zimmermann, "A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 562–585, 2019.
- [2] 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)*, July 2021, pp. 1–6.
- [3] Vignesh V Menon, 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)*. IEEE, 2023, pp. 1673–1678.
- [4] Ioannis Katsavounidis, "Iterative techniques for encoding video content," Feb. 9 2021, US Patent 10,917,644.
- [5] Krishna Srikanth Durbha, Hassene Tmar, Cosmin Stejerean, Ioannis Katsavounidis, and Alan C Bovik, "Bitrate ladder construction using visual information fidelity," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–4.
- [6] Krishna Srikanth Durbha and Alan C Bovik, "Constructing per-shot bitrate ladders using visual information fidelity," *arXiv e-prints*, pp. arXiv-2408.2024.
- [7] 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 *2021 IEEE ICIP*. IEEE, 2021, pp. 2164–2168.
- [8] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara, "Toward A Practical Perceptual Video Quality Metric," *The Netflix Tech Blog*, vol. 6, pp. 2, 2016.
- [9] Rakesh Rao Ramachandra Rao, Steve Göring, Peter List, Werner Robitzka, Bernhard Feiten, Ulf Wüstenhagen, and Alexander Raake, "Bitstream-based model standard for 4k/uhd: Itu-t p. 1204.3—model details, evaluation, analysis and open source implementation," in *2020 QoMEX*. IEEE, 2020.
- [10] Yixu Chen, Zaixi Shang, Hai Wei, Yongjun Wu, and Sriram Sethuraman, "Encoder-quantization-motion-based video quality metrics," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.
- [11] Joshua P. Ebenezer, Zaixi Shang, Yixu Chen, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik, "Hdr or sdr? a subjective and objective study of scaled and compressed videos," 2023.
- [12] Joshua P. Ebenezer, Yixu Chen, Yongjun Wu, Hai Wei, and Sriram Sethuraman, "Subjective and objective quality assessment of high-motion sports videos at low-bitrates," in *2022 IEEE ICIP*, 2022, pp. 521–525.
- [13] Zaixi Shang, Yixu Chen, Yongjun Wu, Hai Wei, and Sriram Sethuraman, "Subjective and objective video quality assessment of high dynamic range sports content," in *Proceedings of the IEEE/CVF WACV Workshops*, January 2023, pp. 556–564.
- [14] Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1–14, 2010.
- [15] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *2010 QoMEX*. IEEE, 2010, pp. 82–87.
- [16] Tobias Hößfeld, 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.
- [17] Zaixi Shang, Joshua P Ebenezer, Abhinav K Venkataraman, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik, "A study of subjective and objective quality assessment of hdr videos," *IEEE Transactions on Image Processing*, vol. 33, pp. 42–57, 2023.
- [18] Maria Perez-Ortiz and Rafal K Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," *arXiv preprint arXiv:1712.03686*, 2017.
- [19] Michela Testolina, Mohsen Jenadeleh, Shima Mohammadi, Shaolin Su, Joao Ascenso, Touradj Ebrahimi, Jon Sneyers, and Dietmar Saupe, "Fine-grained subjective visual quality assessment for high-fidelity compressed images," *arXiv preprint arXiv:2410.09501*, 2024.
- [20] Jing Li, Rafal Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet, "Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation," *Advances in neural information processing systems*, vol. 31, 2018.
- [21] Jing Li, Marcus Barkowsky, and Patrick Le Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3d tv: performances of three different designs," in *Stereoscopic displays and applications XXIV*. SPIE, 2013, pp. 547–558.
- [22] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiastian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al., "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.
- [23] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafal K Mantiuk, "Active sampling for pairwise comparisons via approximate message passing and information gain maximization," in *2020 ICPR*. IEEE, 2021, pp. 2559–2566.
- [24] Shima Mohammadi and Joao Ascenso, "Evaluation of sampling algorithms for a pairwise subjective assessment methodology," in *2022 IEEE ISM*. IEEE, 2022, pp. 288–292.
- [25] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafal K Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2019.
- [26] ITU-R, "Subjective video quality assessment methods for multimedia applications," ITU-R Recommendation Recommendation P.910, 2022.
- [27] Jingwen Zhu, Ali Ak, Patrick Le Callet, Sriram Sethuraman, and Kumar Rahul, "Zrec: Robust recovery of mean and percentile opinion scores," in *2023 IEEE ICIP*. IEEE, 2023, pp. 2630–2634.
- [28] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [29] David J Rogers and Taffee T Tanimoto, "A computer program for classifying plants: The computer is programmed to simulate the taxonomic process of comparing each case with every other case," *Science*, vol. 132, no. 3434, pp. 1115–1118, 1960.
- [30] Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, and Frédéric Dufaux, "Rv-tmo: Large-scale dataset for subjective quality assessment of tone mapped images," *IEEE Transactions on Multimedia*, vol. 25, pp. 6013–6025, 2022.
- [31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [32] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. Ieee, 2003, vol. 2, pp. 1398–1402.