ICCV #5421

BrightRate: Quality Assessment for User-Generated HDR Videos Supplementary Material

Anonymous ICCV submission

Paper ID 5421

Supplementary Material Outline

This supplementary material is organized as follows:

- **Appendix A:** An overview of UGC and HDR video quality assessment challenges.
- **Appendix B.1:** Comprehensive details on video collection, filtering, transcoding, and the bitrate ladder.
- **Appendix B.2:** Full description of our AMT study, including instructions, screening procedure, and rejection criteria.
- **Appendix B.3:** Analysis of inter-subject consistency and SUREAL-based MOS estimation.
- **Appendix C:** Detailed examination of luminance, colorfulness, and spatial-temporal characteristics.
- **Appendix D:** Additional technical specifics on resizing, normalization, and training.
- **Appendix E:** Extended results, ablation studies, and failure case analyses.

A. Background

The explosion of UGC on platforms such as YouTube, Facebook, Instagram, and TikTok has transformed video streaming into a ubiquitous, user-driven experience, generating billions of daily views [1, 12, 13]. However, diverse distortion patterns arising from user editing, compression, and platform-specific processing complicate quality assessment [10, 18]. High Dynamic Range (HDR) imaging, supported by mainstream platforms and devices, offers enhanced visual experiences through a broader luminance and color range. Unlike Standard Dynamic Range (SDR) videos, which are limited to 0.1 to 100 cd/m^2, HDR can represent luminance from 10^{-4} to 10^4 cd/m^2 [8]. HDR10, a widely adopted format, supports 10-bit color depth and Rec. 2020 color gamut (covering 75.8% of the CIE 1931 color space), providing higher peak luminance, improved color accuracy, and more detail in both shadows and highlights, offering a richer, more immersive viewing experience than SDR. The transition to HDR for UGC poses challenges for Video Quality Assessment (VQA) due to increased bit depth, broader luminance range, and complex electro-optical transfer functions (EOTFs) like SMPTE ST 2084 [16]. Traditional SDR-based models fail to capture these HDR-specific features and the variability of distortions from different devices and editing techniques, thereby impeding effective quality prediction. Furthermore, the absence of a publicly available HDR-UGC database has limited the development of HDR-specific VQA models.

B. Details of Dataset Construction

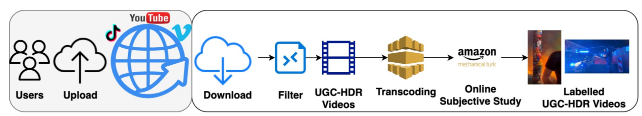


Figure 1. Overview of the dataset preparation approach.

Fig. 1 provides an overview of the entire dataset preparation pipeline. This multi-stage process guarantees that *BrightVQ* reflects authentic HDR-UGC content with diverse distortions

B.1. Video data Collection

Figure 2. HDR specific challenges, and transcoding (on top of ugc) and ugc challenges in *BrightVQ*.

Table 1. Bitladder used for dataset creation. Each video was encoded at multiple bitrates to simulate real-world streaming conditions[1].

Resolution	**Bitrates (Mbps)**
360p	0.2
720p	0.5, 2.0
1080p	0.5, 1.0, 3.0
1080p	Reference

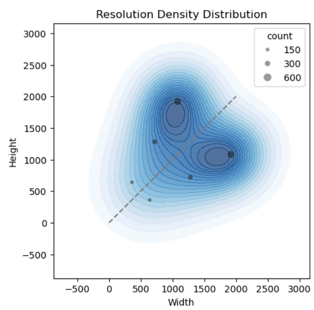


Figure 3. Resolution distribution of *BrightVQ* dataset, maintaining a balanced mix of landscape and portrait videos to study orientation-based perceptual differences.

HDR-UGC videos were collected from Vimeo under Creative Commons licenses to ensure open-source accessibility. An initial pool of over 10,000 videos was automatically filtered using metadata checks for HDR flags, resolution, format consistency, and common categories to remove duplicates and professionally produced content. This was followed by a rigorous manual inspection to verify authentic UGC. Given the nature of UGC, the dataset includes an equal mix of landscape and portrait videos. Fig. 4 shows randomly selected frames from *BrightVQ*, illustrating the diversity in video sizes and aspect ratios. This diversity highlights the broad representation of UGC content in terms of resolution, aspect ratios, and distortions.

Each selected video was truncated to a maximum of 10 seconds using `ffmpeg` [5] and maintained at up to 1080p resolution. To simulate the viewing experience on social media platforms, where videos are often transcoded, we applied a bitrate ladder inspired by industry standards [2, 6] to create the final dataset. Tab. 1 shows the resolution and bitrates used in this bit ladder. The filtered videos were then transcoded following this bitrate ladder to simulate real-world streaming conditions, ensuring a diverse range of compression artifacts. To explore the impact of bitrate selection on perceived video quality, Fig. 6 presents the MOS variations across different bitrate ladders, separately analyzing landscape and portrait videos. The box plot representation highlights the diversity in perceptual quality ratings across different encoding configurations, showing how bitrate and resolution choices affect MOS scores. Fig. 3 illustrates the resolution density distribution of videos in the *BrightVQ* dataset. Fig. 5 further visualizes the compression artifacts introduced through this approach. This multi-stage process guarantees that *BrightVQ* reflects authentic HDR-UGC content with diverse distortions.

[1]Based on YouTube's streaming guidelines [6] and Apple's HLS authoring specifications [2].

B.2. Crowdsourced Subjective Study

We employed Amazon Mechanical Turk (AMT) to collect human quality ratings for our HDR-UGC videos, adapting protocols from previous studies [3, 17]. This is the first large-scale HDR-UGC study conducted on AMT, addressing challenges associated with remote HDR evaluation. To ensure data reliability, we implemented a rigorous multiple-stage filtering process.

The general instruction of this study on AMT is illustrated in Fig. 7. Initially, subjects were presented with detailed instructions and a comprehension quiz (Fig. 8) to confirm their understanding of the rating process. Only those with HDR-capable displays, verified through automated dynamic checks for bit depth, codec support, and display resolution, were allowed to proceed. Before entering the main study, subjects completed a training phase where they rated six HDR videos to familiarize themselves with the interface (Fig. 9). The testing phase followed, in which each participant rated 94 videos using a 0–100 Likert scale (rating method shown in Fig. 10). To ensure rating consistency, we embedded five golden set videos and five duplicate videos within the test set. Ethical considerations are provided in Fig. 11.

To maintain data integrity, we implemented strict rejection criteria at multiple stages:

- **During Instructions:** Participants with incompatible devices were disqualified.
- **During Training:** Continuous HDR and device checks ensured that participants did not switch displays mid-task. Those with incomplete downloads or playback manipulations were excluded.
- **During Testing:** Participants were monitored at 25%, 50%, and 75% of task completion. Those exhibiting over 50% playback issues or inconsistent ratings on duplicate/golden set videos (deviations exceeding 20–25%) were removed.

In total, over 200 subjects provided 73,794 ratings (an average of 35 ratings per video).

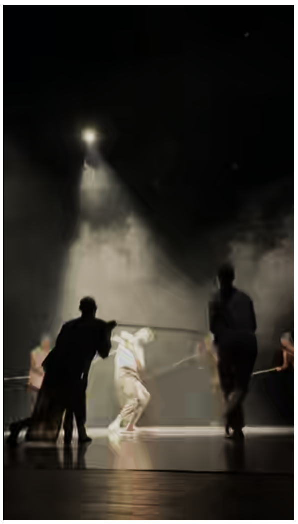
B.3. Subjective Score Processing

To evaluate inter-subject consistency, we randomly split all MOS ratings into two independent groups and computed the Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) between them. As shown in Fig. 12a, the study achieved a median SRCC of

(a) Sports (MOS: 31.77)
[Video Link]

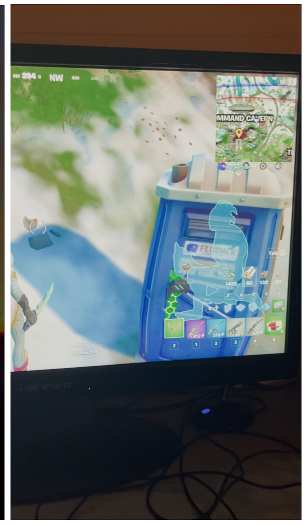
(b) Cats (MOS: 41.61)
[Video Link]

(c) Christmas Tree (MOS: 64.95)
[Video Link]



(d) Dance (MOS: 37.05)
[Video Link]

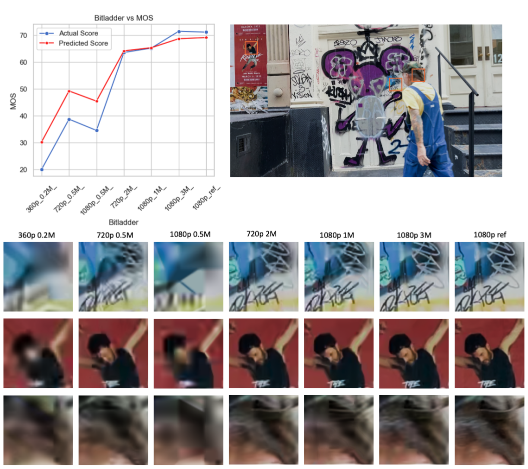
(e) Art (MOS: 50.89)
[Video Link]



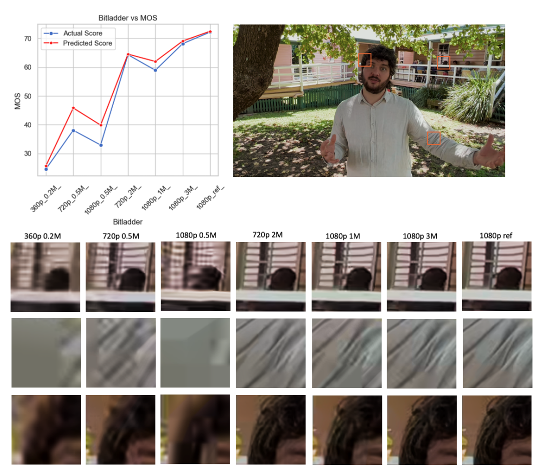
(f) Game (MOS: 65.02)
[Video Link]

(g) Car (MOS: 71.06)
[Video Link]

Figure 4. Example frames from *BrightVQ* dataset. Each frame is presented with its category, the MOS for the video and a direct video access link.

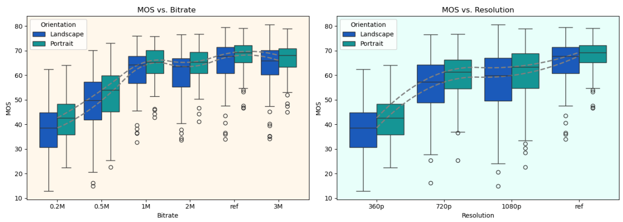


(a)



(b)

Figure 5. The combination of MOS vs. predicted score plots with visual comparisons of specific image regions to highlight the correlation between distortion and MOS across different bitrates and resolutions.



(a) MOS vs. Bitrate. (b) MOS vs. Resolution.

Figure 6. (a) and (b) show the MOS variations across bitrate and resolution respectively for *BrightVQ*.

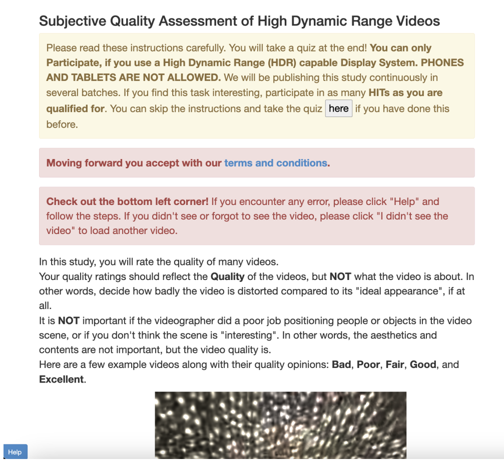


Figure 7. General instruction of this study on AMT.

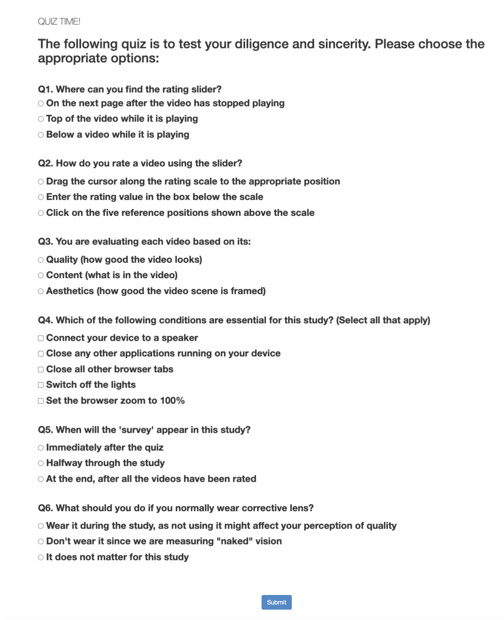


Figure 8. Quiz phase on AMT.

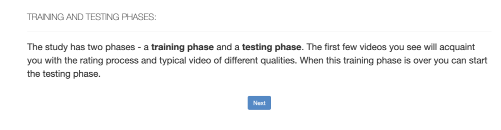


Figure 9. Train-test Instruction phase on AMT.

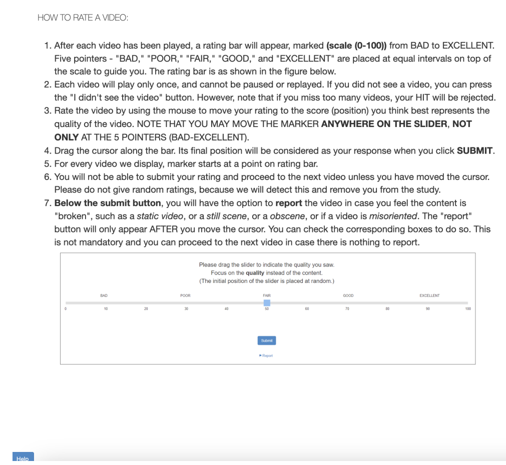


Figure 10. Rating instructions on AMT.

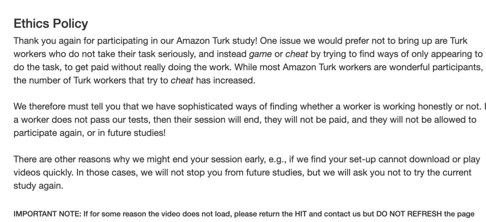
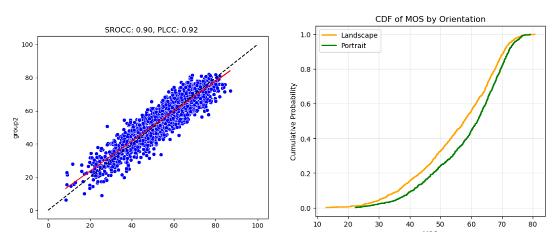


Figure 11. Ethics policy on AMT.



(a) (b)

Figure 12. (a). Inter-subject correlation. (b). MOS CDF distribution of all videos in *BrightVQ*.

0.90 and a median PLCC of 0.92, demonstrating a high level of agreement between independent participant groups. This strong correlation validates the effectiveness of our data collection methodology, which incorporated device filtering, training phases, and golden set validation to ensure consistent and reliable subjective ratings.

We computed Mean Opinion Scores (MOS) using the SUREAL method [9], which refines traditional MOS computation by accounting for individual subject bias and rating inconsistency. Traditional MOS calculations typically employ a hard rejection approach, where raters failing predefined consistency criteria (e.g., ITU-R BT.500-14 outlier detection [7]) are completely excluded from the analysis. However, this method discards potentially useful data and does not account for varying levels of rating reliability among retained subjects. SUREAL takes a soft rejection approach by modeling each rating probabilistically. Each rating S_{ij} from subject i for video j is modeled as:

$$S_{ij} = \psi_j + \Delta_i + \nu_i X, \quad X \sim \mathcal{N}(0,1), \tag{1}$$

where ψ_j represents the true quality of video j, Δ_i captures the bias of subject i, and ν_i reflects the rating inconsistency of subject i. The parameters are estimated using Maximum Likelihood Estimation (MLE), maximizing the likelihood. Unlike hard rejection, which entirely removes outliers, SUREAL downweights ratings from less reliable subjects. This ensures that MOS values reflect true perceptual quality while mitigating distortions from inconsistent raters. By applying SUREAL, we obtained more stable MOS estimates, which accurately reflect the perceptual quality of HDR content across diverse video conditions. The CDF distribution of all videos in *BrightVQ* are shown in Fig. 12b.

C. Analysis of the HDR content

In this section, we provide a detailed analysis of the dataset's key characteristics, focusing on luminance and color distribution, spatial-temporal diversity, perceptual quality trends, and HDR-specific challenges.

Fig. 13 presents the distribution of luma and colorfulness across the 300 source HDR videos in BrightQA. The first three histograms illustrate the minimum, maximum, and mean luma values, highlighting the variation in brightness levels across different videos. This demonstrates that the dataset includes both dark and bright HDR scenes, ensuring a wide dynamic range. The fourth histogram shows the colorfulness distribution, reflecting variations in chromatic intensity across different videos.

To further quantify the diversity in content complexity, Fig. 14 presents an analysis of spatial-temporal complexity, spatial information (SI), and temporal information (TI) within the dataset. The scatter plots in Fig. 14 (a)-(c) demonstrate the variability in SI and TI values, showing a wide distribution of motion and texture complexity across the dataset. Higher SI values typically correspond to detailed textures and sharp edges, while higher TI values indicate rapid motion or dynamic scenes. The dataset covers both high-detail static scenes and fast-moving dynamic content, ensuring its suitability for evaluating compression and HDR features across different motion characteristics.

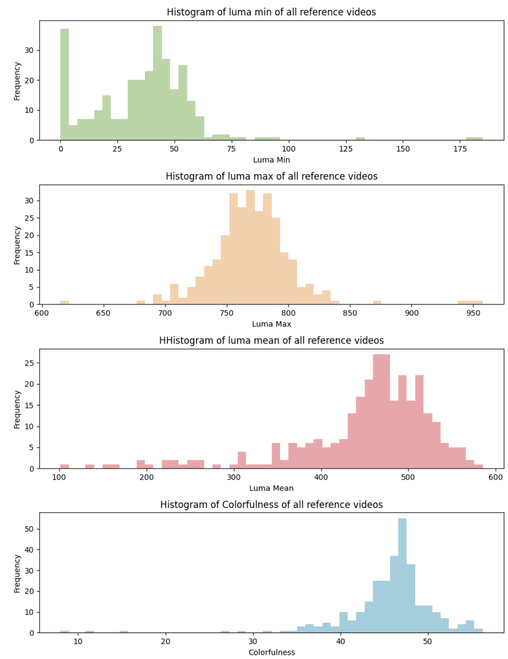
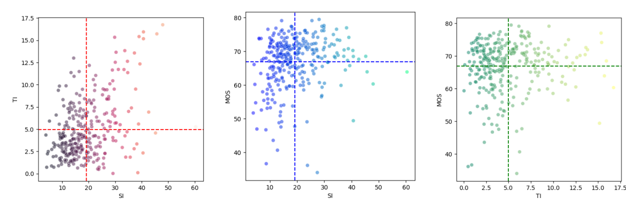


Figure 13. Distribution of luma and colorfulness of the source HDR-UGC videos in *BrightVQ* dataset.



(a) SITI (b) MOS vs. SI (c) MOS vs. TI

Figure 14. (a) Spatial-Temporal Complexity, (b) MOS vs. Spatial Information (SI), and (c) MOS vs. Temporal Information (TI).

Fig. 15 demonstrates the diversity of BrightVQ dataset in both aesthetic and technical aspects. The scatter plot shows a wide range of ratings, with each point representing a video and color-coded by its actual subjective quality score. The marginal histograms further highlight the distribution of scores, illustrating the broad variation in perceptual and technical quality across different content. The BrightVQ dataset presents a diverse range of HDR-UGC content, covering natural landscapes, indoor scenes, and various complex lighting conditions, capturing both HDR-specific and UGC-specific distortions. This diversity ensures that BrightVQ provides a comprehensive and realistic benchmark for evaluating video quality.

Figure 16 provides examples of HDR videos subjected to various spatial resolutions and bitrates, along with their MOS and luminance histograms. In (a) and (c), the higher-resolution, higher-bitrate frames retain more de-

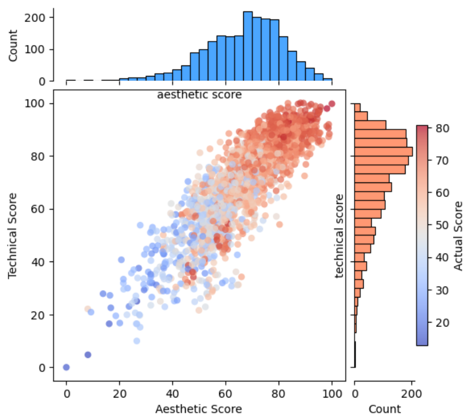


Figure 15. Diversity in aesthetic and technical quality scores in *BrightVQ* datset.

tails and exhibit fewer artifacts, whereas lower-resolution, lower-bitrate versions show noticeable blocking and banding—particularly in the extreme luma regions. Subfigures (b) and (d) illustrate the broader luminance distribution characteristic of HDR, indicating significant content in both low and high intensity ranges. Such distributions underscore the importance of HDR-specific processing, since distortions in extreme luminance regions can disproportionately affect subjective quality.

D. More details on Implementation

Here we detail the key implementation steps of our *BrightRate* model. For semantic features, input frames are resized to 224×224 and passed through the CLIP image encoder (ViT-B32) [14], yielding high-level semantic representations. UGC features are extracted using CONTRIQUE [11] at two scales: the original frame and a half-resolution version following the original implementation [11]. For HDR features, we convert each frame to YUV, extract the luminance channel $\mathbf{Y}^t \in [0, 1]$, and apply a piecewise expansive non-linearity over a 31×31 window with $\beta = 4$ [4, 15]. We then compute MSCN coefficients on the transformed luminance and model their statistics using GGD/AGGD to obtain HDR features $\mathcal{H}^t$. Temporal dynamics are captured by computing the absolute differences between consecutive CONTRIQUE [11] features:

$$\Delta \mathcal{U}^t = \left| \mathcal{U}^t - \mathcal{U}^{t-1} \right|, \tag{2}$$

which are then globally averaged and concatenated with the spatial features. The final clip-level representation is formed by normalizing and concatenating the UGC, semantic, and HDR features:

$$\mathbf{z} = \mathrm{Norm}\left(\overline{\mathcal{U}} \oplus \overline{\mathcal{E}} \oplus \overline{\mathcal{H}}\right). \tag{3}$$

This vector is then fed to a Support Vector Regressor (SVR) to predict the MOS. We train the SVR using 5-fold cross-validation to optimize the regularization parameter and repeat the process over 100 random splits, reporting the median performance. The training loss is given by:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(Q_i - \hat{Q}_i \right)^2, \tag{4}$$

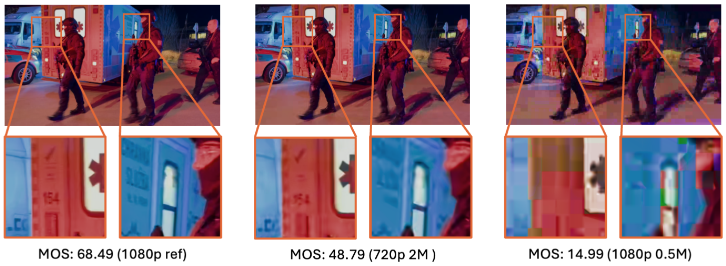
where Q_i and $\hat{Q}_i$ denote the ground-truth and predicted MOS, respectively. Only the regressor is trained, while the feature extraction modules remain fixed. These implementation details ensure a robust and efficient extraction of multi-scale UGC, semantic, and HDR features, enabling accurate quality prediction on HDR-UGC videos.

E. More Experimental Results

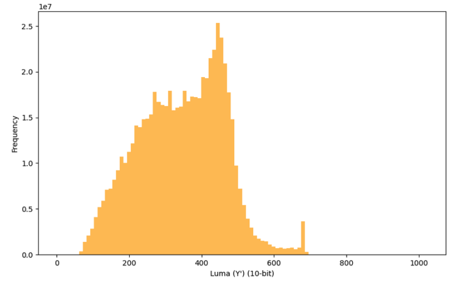
To assess the effectiveness of existing No-Reference Video Quality Assessment models on the *BrightVQ* dataset, we conducted a comprehensive evaluation of multiple state-of-the-art methods. Fig. 17 expands upon Fig. 9, which presented results for only six models, by providing a more extensive comparison across 13 NR-VQA model. The scatter plots compare predicted scores vs. MOS, with red parametric fitting lines highlighting the correlation trends, while the Pearson correlation coefficient (r) quantifies each model's predictive performance.

Among the evaluated models, some traditional handcrafted feature-based approaches exhibit limited correlation with MOS, highlighting their challenges in capturing the complexity of HDR-specific distortions in UGC content. Deep learning-based methods show stronger performance, with several achieving a higher degree of correlation by leveraging learned features and spatiotemporal representations. Moreover, HDR-specific VQA models generally outperform generic NR-VQA methods, demonstrating the importance of HDR-aware architectures in perceptual quality assessment. Our proposed *BrightRate* model achieves the highest correlation (r = 0.91), significantly outperforming other approaches. The scatter plot for *BrightRate* shows a strong linear relationship between predicted scores and MOS, indicating its high accuracy and reliability in evaluating HDR video quality.

Fig. 18 presents several failure cases where the predicted video quality scores deviate significantly from the actual MOS. These discrepancies highlight limitations in the model's ability to accurately predict perceptual quality under certain conditions. Fig. 18 (a) illustrates cases where low-resolution, highly compressed videos received higher-than-expected predictions. The close-up patches of compressed video artifacts reveal that blockiness and blurring effects are not always adequately penalized by the model,

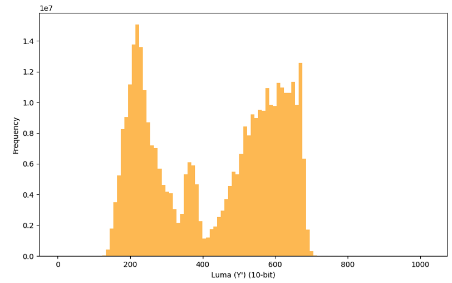


(a)



(b)

(c)



(d)

Figure 16. Illustrations of HDR content under different resolutions and bitrates. (a) and (c) show reference frames at 1080p and their lower-resolution, lower-bitrate counterparts, with red boxes highlighting high luma areas with artifacts (e.g., blocking, color banding) become more pronounced. The corresponding MOS values indicate how these distortions affect subjective perception. (b) and (d) present the luminance histograms of the respective frames, revealing a broader distribution for HDR content that spans both low and high luminance ranges. This demonstrates the increased complexity of HDR videos.

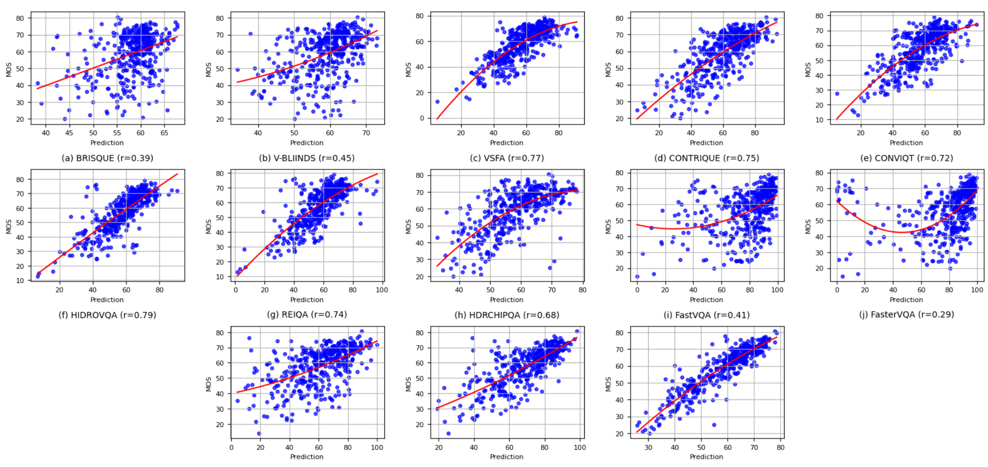
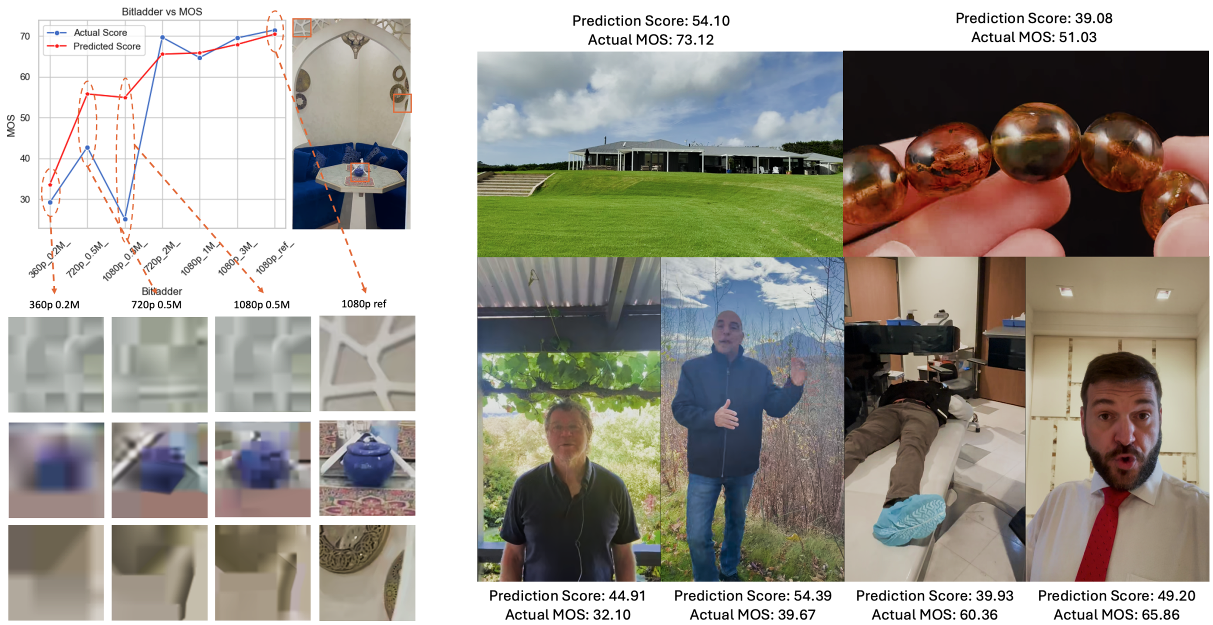


Figure 17. Scatter plots of actual MOS vs. predicted scores for 13 methods evaluated on *BrightVQ*, with parametric fits $l(s)$ in red. A tighter clustering around the diagonal curve indicates a stronger alignment with subjective opinions. Methods yielding narrower scatter demonstrate higher predictive accuracy and consistency, underscoring their ability to capture the underlying perceptual quality cues.

leading to overestimated quality scores in severely compressed videos. Fig. 18 (b) show video screenshots with complex textures, reflections, or dynamic lighting, where the model struggles to properly assess fine details and HDR characteristics. In videos with human subjects, facial expressions, lighting conditions, or background complexity may lead to misinterpretations of perceptual quality by the model. These failure cases highlight the need for further re-

ICCV
#5421



(a) (b)

Figure 18. Failure cases in *BrightRate* predictions.

finement in *BrightRate*'s HDR-aware feature extraction and compression robustness, ensuring improved alignment with human perception.

References

[1] 99Firms. Facebook video statistics, 2024. [Online].

[2] Apple Inc. Hls authoring specification for apple devices, 2024. Accessed: Feb. 2024.

[3] Yu-Chih Chen, Avinab Saha, Alexandre Chapiro, Christian Häne, Jean-Charles Bazin, Bo Qiu, Stefano Zanetti, Ioannis Katsavounidis, and Alan C. Bovik. Subjective and objective quality assessment of rendered human avatar videos in virtual reality. *IEEE Transactions on Image Processing*, 33: 5740–5754, 2024.

[4] Joshua P. Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Hdr-chipqa: No-reference quality assessment on high dynamic range videos, 2023.

[5] FFmpeg Developers. Ffmpeg. https://ffmpeg.org/. Accessed: 2025-02-04.

[6] Google Support. Recommended upload encoding settings, 2024. Accessed: Feb. 2024.

[7] International Telecommunication Union. Methodology for the Subjective Assessment of the Quality of Television Pictures. Technical Report BT.500-14, International Telecommunication Union, 2019.

[8] ITU. Bt.2020 : Parameter values for ultra-high definition television systems for production and international programme exchange,. https://glenwing.github.io/docs/ITU-R-BT.2020-1.pdf.

[9] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments. In *Electronic Imaging*, pages 131–1–131–14, 2020.

[10] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos, 2024.

[11] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Trans. Image Process.*, 31: 4149–4161, 2022.

[12] Maryam Mohsin. 10 youtube statistics every marketer should know in 2020, 2020. [Online].

[13] Omnicore. Tiktok by the numbers, 2024. [Online].

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[15] Zaixi Shang, Joshua P Ebenezer, Abhinau K Venkataramanan, 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*, 33:42–57, 2023.

[16] SMPTE. High dynamic range electrooptical transfer function of mastering reference displays. https://pub.smpte.org/latest/st2084/st2084-2014.pdf.

[17] Abhinau K. Venkataramanan and Alan C. Bovik. Subjective quality assessment of compressed tone-mapped high dynamic range videos, 2024.

[18] Zicheng Zhang, Wei Wu, Wei Sun, Dangyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos, 2023.