CHUG: CROWDSOURCED USER-GENERATED HDR VIDEO QUALITY DATASET

Shreshth Saini, Alan C. Bovik

The University of Texas at Austin
Austin, TX, USA

Neil Birkbeck, Yilin Wang, Balu Adsumilli

YouTube, Google Inc.
Mountain View, CA, USA

ABSTRACT

High Dynamic Range (HDR) videos enhance visual experiences with superior brightness, contrast, and color depth. The surge of User-Generated Content (UGC) on platforms like YouTube and TikTok introduces unique challenges for HDR video quality assessment (VQA) due to diverse capture conditions, editing artifacts, and compression distortions. Existing HDR-VQA datasets primarily focus on professionally generated content (PGC), leaving a gap in understanding real-world UGC-HDR degradations. To address this, we introduce **CHUG**: Crowdsourced User-Generated HDR Video Quality Dataset, the first large-scale subjective study on UGC-HDR quality. CHUG comprises 856 UGC-HDR source videos, transcoded across multiple resolutions and bitrates to simulate real-world scenarios, totaling 5,992 videos. A large-scale study via Amazon Mechanical Turk collected 211,848 perceptual ratings. CHUG provides a benchmark for analyzing UGC-specific distortions in HDR videos. We anticipate CHUG will advance No-Reference (NR) HDR-VQA research by offering a large-scale, diverse, and real-world UGC dataset. The dataset is publicly available at: https://shreshthsaini.github.io/CHUG/.

Index Terms— Crowdsourced, High Dynamic Range (HDR), Video Quality Assessment, HDR VQA Dataset, User-Generated Content (UGC)

1. INTRODUCTION

The rapid growth of User-Generated Content (UGC) on platforms such as YouTube, Instagram, and TikTok has transformed digital media consumption [1, 2, 3]. UGC is characterized by diverse capture conditions, user-editing effects, and platform-specific compression, which complicates Video Quality Assessment (VQA) [4, 5]. Simultaneously, High Dynamic Range (HDR) is gaining adoption due to its wider color gamut, higher bit depth, and enhanced luminance, enabling improved perceptual quality [6, 7]. However, UGC-HDR VQA remains an open problem due to HDR-specific distortions such as banding, tone-mapping artifacts, and exposure non-uniformity [8, 9]. Existing HDR datasets such as LIVE-HDR [10] and SFV+HDR [11] focus on professionally generated content (PGC) and lack scale and diversity to represent real-world UGC scenarios. LIVE-HDR [10] contains only 31 reference HDR videos with professionally captured content. The ITM-HDR-VQA dataset [12] offers 200 inverse tone-mapped HDR videos, while the KVQ dataset [13] and SFV+HDR dataset [11] focus on portrait short-form content. To the best of our knowledge, there exists no large-scale, publicly available UGC-HDR dataset that is representative of real-world HDR distortions. CHUG addresses these gaps by providing the largest UGC-HDR dataset with real-world distortions, authentic content diversity, and high-quality subjective scores, serving as a benchmark for advancing NR-VQA models. The key characteristics of this dataset are summarized in Table 1. The key contributions are:

- CHUG includes 856 UGC-HDR source videos, transcoded across resolutions and bitrates, creating 5,992 videos—the largest HDR and UGC-HDR dataset to date.
- CHUG applies a bitrate ladder encoding strategy, replicating social media compression on UGC-HDR videos.
- The first large-scale HDR study on Amazon Mechanical Turk (AMT), collecting 211,848 ratings, with each video rated by 35 subjects on average.
- CHUG serves as a benchmark for UGC-HDR distortions, aiding NR-VQA model development for real-world HDR streaming.

Table 1: Overview of the CHUG Dataset. The dataset comprises diverse UGC-HDR videos across multiple resolutions and bitrates, with extensive subjective quality annotations.

Attribute	**Details**
Video Specifications	
Format	Rec. 2020, 10-bit, PQ
Resolutions	1920×1080, 1080×1920, 1280×720, 720×1280, 640×360, 360×640
Duration	4 - 10 sec.
Dataset Statistics	
Reference Videos	856 (428 Portrait, 428 Landscape)
Total Videos	5,992
Total Scores	211,848
Avg. Scores per Video	35

Fig. 1: Sample frames from the CHUG dataset, showcasing diverse real-world UGC-HDR content with variations in lighting, motion, orientation, and distortions. Best viewed when zoomed in.

2. CROWDSOURCING UGC-HDR VIDEOS

In this section, we discuss the construction of a large-scale UGC-HDR video quality dataset by collecting real-world HDR videos from users. To ensure diverse content, orientations, and distortions, we curated the dataset and applied bitrate ladder encoding to simulate real-world streaming conditions.

2.1. Source Video Collection and UGC Diversity

The dataset was built through an open call for UGC-HDR video submissions, where contributors uploaded HDR videos from personal devices such as iPhones, Samsung Galaxy, Google Pixel, and OnePlus smartphones. All submissions included appropriate consent and rights transfer agreements. We applied strict filtering criteria to all submitted videos to ensure dataset integrity by removing duplicates, objectionable content, and static or minimally dynamic videos. Each video was trimmed to a maximum 10 seconds using `ffmpeg` [14], with no resolution upscaling or downscaling beyond 1080p to preserve original quality.

Table 2: Bitrate ladder 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

An equal mix of landscape and portrait videos was maintained to analyze orientation-based quality perception. Fig. 2 presents the resolution distribution of CHUG, although different resolution has different number of videos but for single resolution there is an equal split of landscape and portrait videos. The dataset spans urban environments, natural landscapes, indoor vlogs, and sports recordings, covering a range of lighting conditions, including daylight, nighttime, and extreme brightness scenarios, see Fig. 1.

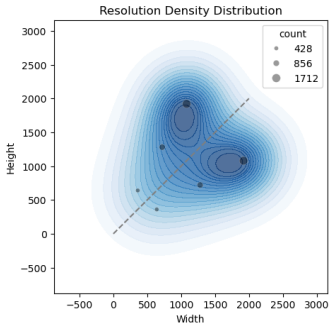


Fig. 2: Resolution distribution of CHUG dataset, maintaining a balanced mix of landscape and portrait videos to study orientation-based perceptual differences.

2.2. Bitrate Ladder for Realistic Streaming Simulation

UGC videos uploaded to social media platforms undergo platform-specific compression and transcoding, introducing bitrate-dependent distortions. To replicate these effects, we applied a bitladder encoding strategy [10], following YouTube's streaming guidelines [15] and Apple's HLS authoring specifications [16]. Table 2 details the bitladder used in CHUG, ensuring that each video was encoded at multiple bitrates to simulate real-world streaming conditions.

This encoding process introduced controlled compression artifacts, allowing us to analyze how bitrate variation impacts subjective HDR video quality. Fig. 3 illustrates visible compression artifacts at different resolutions. The leftmost frame represents the original 1080p reference video, the middle frame is 720p at 2 Mbps, and the rightmost frame is 360p at 0.2 Mbps, showing noticeable degradation due to aggressive compression. The combination of authentic UGC-HDR content and real-world streaming simulation makes CHUG a challenging benchmark dataset.

[1] Based on YouTube's streaming guidelines [15] and Apple's HLS authoring specifications [16].

Table 3: Comparison of CHUG with existing HDR VQA datasets.

Dataset	Format	Total Videos(Ref.)	Source	Total opinions	Orientation
LIVE-HDR	Rec. 2020, HDR10, PGC	310 (31)	Internet Archive	20,400	Landscape
SFV+HDR(*only HDR*)	Rec. 2020, HDR10, UGC	300	YouTube	N/A	Portrait
CHUG (Ours)	Rec. 2020, HDR10, UGC	5,992 (856)	Crowdsourced	211,848	Portrait/Landscape

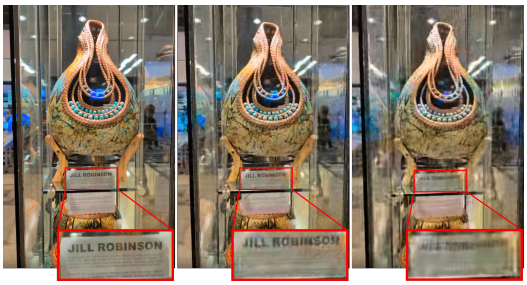


Fig. 3: Compression artifacts introduced via bitladder. Left: 1080p reference, Middle: 720p at 2 Mbps, Right: 360p at 0.2 Mbps.

3. DETAILS OF SUBJECTIVE STUDY

In this section, we describe the subjective study conducted using Amazon Mechanical Turk (AMT) [17] to gather perceptual opinion scores for UGC-HDR videos. This is the first large-scale AMT-based UGC-HDR study, overcoming remote HDR evaluation challenges such as device compatibility, display limitations, and network constraints. To ensure robust data collection, strict filtering criteria were implemented, resulting in 211,848 ratings from 700+ subjects, with each video receiving an average of 35 ratings.

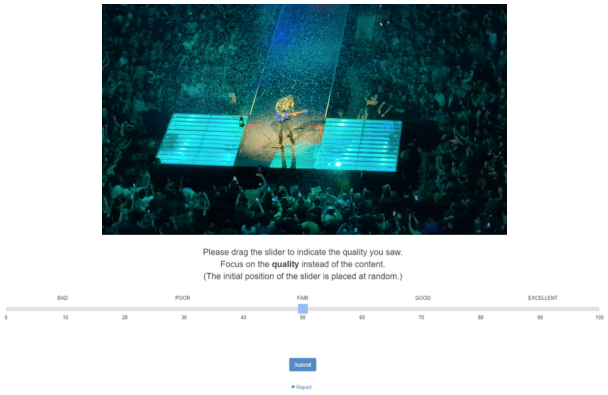


Fig. 4: Rating interface used for the AMT study (Best viewed zoomed in).

3.1. Study Design and Rating Procedure

Each Human Intelligence Task (HIT) pipeline began with instructions and a brief quiz to confirm participant understanding. Subjects then completed a training phase by rating six HDR videos, familiarizing themselves with the rating interface , see Fig. 4. The testing phase involved rating 94 videos, each presented only once, on a 0-100 Likert scale. Videos were pre-loaded to prevent buffering, and participants could flag videos for technical or content issues. At the end of each session, subjects provided device specifications, demographics, and viewing conditions. A pilot study was conducted with experienced AMT users to establish a golden set of 1428 videos, later used for participant reliability filtering.

3.2. Participant Screening and Quality Control

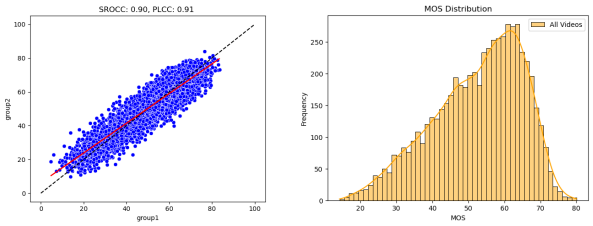
To ensure data reliability, multiple subject and session rejection criteria were applied [4, 18], categorized as:
Pre-screening: Participants with incompatible devices (e.g., non-HDR displays) were blocked from proceeding. We dynamically checked bit depth, HEVC codec support, display resolution, and network speed.
Training Phase: Continuous HDR validation was performed to detect any device setting changes mid-task. Playback completion tracking was enforced to prevent participants from skipping through videos.
Testing Phase: Progressive quality checks were applied at 25%, 50%, and 75% task completion. Subjects were removed if $>$50% of their ratings were flagged for playback issues or if they exhibited inconsistent rating behavior.
Post-study validation: Of the 94 test videos, 10 were control videos (5 duplicates+5 golden set). Subjects with deviations exceeding 20-25% on repeated/golden videos were excluded.

After subject rejection, a rigorous data cleaning process was implemented to ensure the reliability of collected ratings. First, all ratings from disqualified subjects or those experiencing over 50% playback issues were removed. Second, responses from participants who reported not wearing necessary vision correction were excluded. Finally, we applied ITU-R BT.500-14 [19] filtering criteria, leading to the removal of 60 additional subjects. After these refinements, the cleaned dataset contained 211,848 high-quality ratings, ensuring robust subjective evaluation.

4. DATA ANALYSIS

In this section, we analyze the collected subjective scores to assess the quality distribution, inter-subject agreement, and variations in MOS based on video properties such as resolution, bitrate, orientation, and content complexity. We further

compare CHUG against LIVE-HDR and SFV+HDR, highlighting its diversity and representation of real-world UGC-HDR quality.



(a) Inter-subject correlation (b) MOS distribution

Fig. 5: (a) MOS distribution of all videos; (b) Inter-subject correlation

4.1. Processing of Subjective Scores

To obtain reliable Mean Opinion Scores (MOS), we employed the SUREAL (Subjective Reliability) method [20], a robust statistical approach that accounts for subject biases and inconsistencies. Unlike traditional MOS calculations, which simply average subjective ratings [21], SUREAL computes a Maximum Likelihood Estimate (MLE) of the true video quality, making it robust to outliers and unreliable subjects. The opinion scores S_{i_j} given by subject i for video j were modeled as:

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

Here, ψ_j represents the true quality of video j. Δ_i accounts for the rating bias of subject i. ν_i models the inconsistency of subject i, ensuring that subjects with erratic rating behavior have less influence. The parameters ψ_j, Δ_i, and ν_i were estimated using the Newton-Raphson optimization method to maximize the log-likelihood of observed scores.

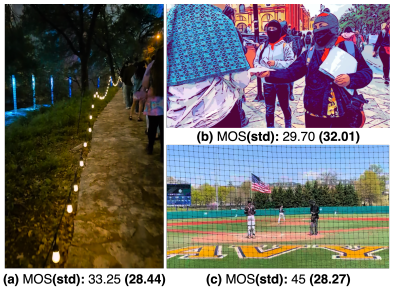
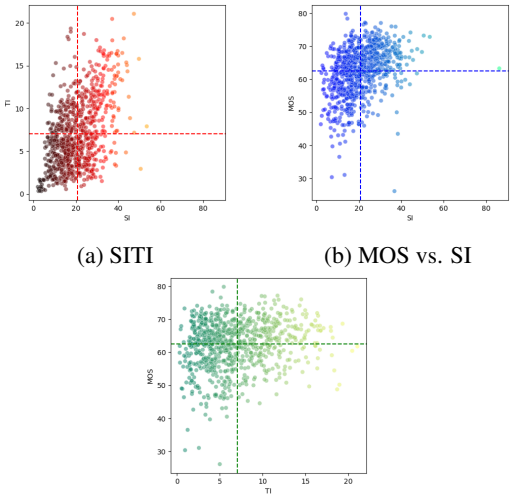


Fig. 6: Examples of challenging videos with high MOS standard deviation. Night-time scenes, extreme filters, and occlusions caused rating inconsistencies.

4.2. Challenging Content and High Variance MOS

Certain videos in CHUG exhibited high MOS variance, reflecting significant subjective disagreement among raters. Fig. 6 highlights examples of such challenging content, where night scenes, extreme filters, obstructions, and complex lighting conditions led to inconsistent ratings. These variations often result in a higher standard deviation in MOS, making them difficult for both human evaluators and objective quality assessment models. A major strength of CHUG is its inclusion of such real-world challenging scenarios, which are often missing from existing HDR datasets like LIVE-HDR [10], where professionally captured videos tend to have uniform lighting, controlled exposure, and minimal distortions. In contrast, CHUG incorporates UGC-HDR content from diverse real-world settings, making it a more versatile and realistic benchmark for NR-VQA models.



(a) SITI (b) MOS vs. SI

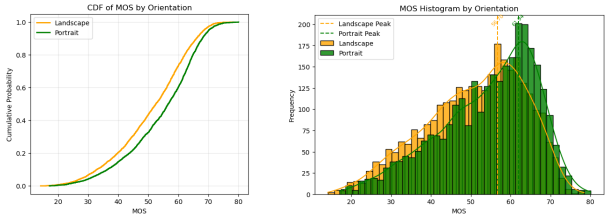
(c) MOS vs. TI

Fig. 7: (a) Spatial-Temporal Complexity, (b) MOS vs. Spatial Information (SI), and (c) MOS vs. Temporal Information (TI).

4.3. Spatial-Temporal Features vs. MOS

We analyzed the relationship between spatial (SI) and temporal (TI) complexity and perceived video quality. Fig. 7a visualizes the joint distribution of SI and TI across the dataset. Most videos cluster in the mid-to-high SI range with low-to-moderate TI, aligning with common UGC-HDR content characteristics. The spread of SI and TI values highlights CHUG's diversity in motion and texture complexity. Fig. 7b indicates a positive correlation between SI and MOS, suggesting that videos with higher spatial details, such as textures and sharp edges, tend to receive higher quality ratings. This trend aligns with perceptual expectations.However, we also observe that extremely high SI values do not necessarily lead to the

highest MOS, possibly due to compression artifacts becoming more noticeable in highly detailed regions. Fig. 7c shows a non-monotonic trend between TI and MOS. While moderate motion complexity is associated with higher MOS, excessive motion (high TI) often results in lower ratings. This degradation is primarily due to motion compression artifacts, where aggressive encoding leads to blurring, ghosting, or unnatural frame interpolation effects.

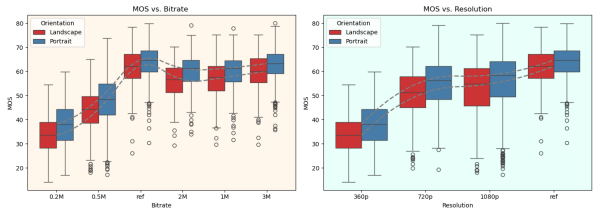


(a) MOS CDF for orientations (b) MOS distribution by orientation

Fig. 8: MOS analysis for landscape vs. portrait videos.

4.4. Effect of Orientation on MOS

Figure 8 analyzes the impact of video orientation (landscape vs. portrait) on perceived quality. Fig. 8b shows a high degree of overlap between landscape and portrait MOS distributions, indicating that orientation alone does not significantly influence perceptual quality. Both orientations exhibit a similar MOS range and distribution shape, suggesting that factors such as content type, motion, and compression artifacts play a more dominant role in quality perception than orientation. Fig. 8a provides a finer comparison using CDF curves. The curves indicate that portrait videos tend to receive slightly higher MOS scores, particularly in the mid-to-high quality range.



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

Fig. 9: MOS variations across bitrate, resolution, and combined bitladder.

4.5. MOS Across Resolution and Bitrate

Figure 9 illustrates how resolution and bitrate influence perceptual quality. Fig. 9a shows that MOS generally increases with bitrate, confirming that higher bitrates preserve visual quality by reducing compression artifacts. However, low-bitrate videos (0.2M, 0.5M) exhibit significant quality degradation, with MOS values spread across a wider range, indicating variability in perceptual impact depending on content complexity. Portrait videos tend to have slightly higher MOS across bitrate range. Fig. 9b highlights the direct correlation between resolution and perceived quality. As expected 360p videos receive the lowest MOS, while MOS steadily improves at 720p, and 1080p. The overlapping MOS distributions for 720p and 1080p suggest diminishing perceptual gains beyond a certain resolution threshold, influenced by content type and display scaling effects.

4.6. Comparing CHUG with Existing HDR Datasets

Figure 10 presents a comparison of MOS distributions across CHUG, LIVE-HDR, and SFV+HDR on common scale. LIVE-HDR[10] and SFV+HDR[11] exhibit a strong skew towards high MOS values, indicating a lack of diverse quality variations. These datasets primarily contain professionally generated or high-quality UGC, limiting their applicability in real-world HDR-VQA tasks. CHUG demonstrates a broader MOS distribution, covering low, medium, and high-quality HDR videos. This highlights its effectiveness in capturing diverse distortions and realistic UGC-HDR variations seen in modern streaming platforms. These findings reinforce the need for datasets like CHUG to bridge the gap between professional HDR content and real-world UGC-HDR challenges.

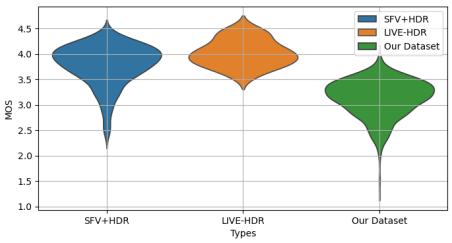


Fig. 10: Comparison of MOS distributions: CHUG vs. LIVE-HDR vs. SFV+HDR.

5. CONCLUSION

In this paper, we introduced CHUG, a large-scale UGC-HDR video quality dataset with 5,992 videos and 211,848 subjective ratings collected via Amazon Mechanical Turk (AMT). Using SUREAL-based MOS computation, we ensured robust quality estimates. Our analysis highlights spatial-temporal complexity, orientation effects, and bitrate-resolution trade-offs in UGC-HDR videos. CHUG serves as a benchmark for No-Reference HDR-VQA models, reflecting real-world social media HDR content. The dataset and scores will be publicly available upon publication.

6. REFERENCES

[1] Omnicore, "Tiktok by the numbers," 2024, [Online].

[2] 99Firms, "Facebook video statistics," 2024, [Online].

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

[4] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik, "Patch-vq: 'patching up' the video quality problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14019–14029.

[5] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[6] Consumer Technology Association (CTA), "Television technology consumer definitions," 2024, [Online; accessed February 2024].

[7] International Telecommunication Union, "BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange," Tech. Rep., International Telecommunication Union, 2018.

[8] 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," *arXiv preprint arXiv:2304.13156*, 2023.

[9] Shreshth Saini, Avinab Saha, and Alan C. Bovik, "HIDRO-VQA: High dynamic range oracle for video quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024, pp. 469–479.

[10] Zaixi Shang, Joshua P. Ebenezer, Alan C. Bovik, Yilin Wu, Hong Wei, and Sethuraman Seshadri, "Subjective assessment of high dynamic range videos under different ambient conditions," in *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 2301–2305, IEEE.

[11] Yilin Wang, Joong Gon Yim, Neil Birkbeck, and Balu Adsumilli, "Youtube sfv+hdr quality dataset," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024, pp. 96–102.

[12] Fei Zhou, Shuhong Yuan, Zhijie Liang, and Guoping Qiu, "Itm-hdr-vqa dataset," 2023.

[13] 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," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 25963–25973.

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

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

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

[17] Amazon Mechanical Turk, "Amazon mechanical turk: Artificial artificial intelligence," https://www.mturk.com, Accessed: 2025-02-05.

[18] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Tech. Rep. BT.500-14, International Telecommunication Union, 2019.

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

[21] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Tech. Rep. BT.500-11, International Telecommunication Union, 2002.