This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Seeing Beyond 8bits: Subjective and Objective Quality Assessment of HDR-UGC Videos

Shreshth Saini[1]* Bowen Chen[1] Yilin Wang[2] Neil Birkbeck[2]
Balu Adsumilli[2] Alan C. Bovik[1]

[1]Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin
[2]Google Inc., Mountain View, CA, USA

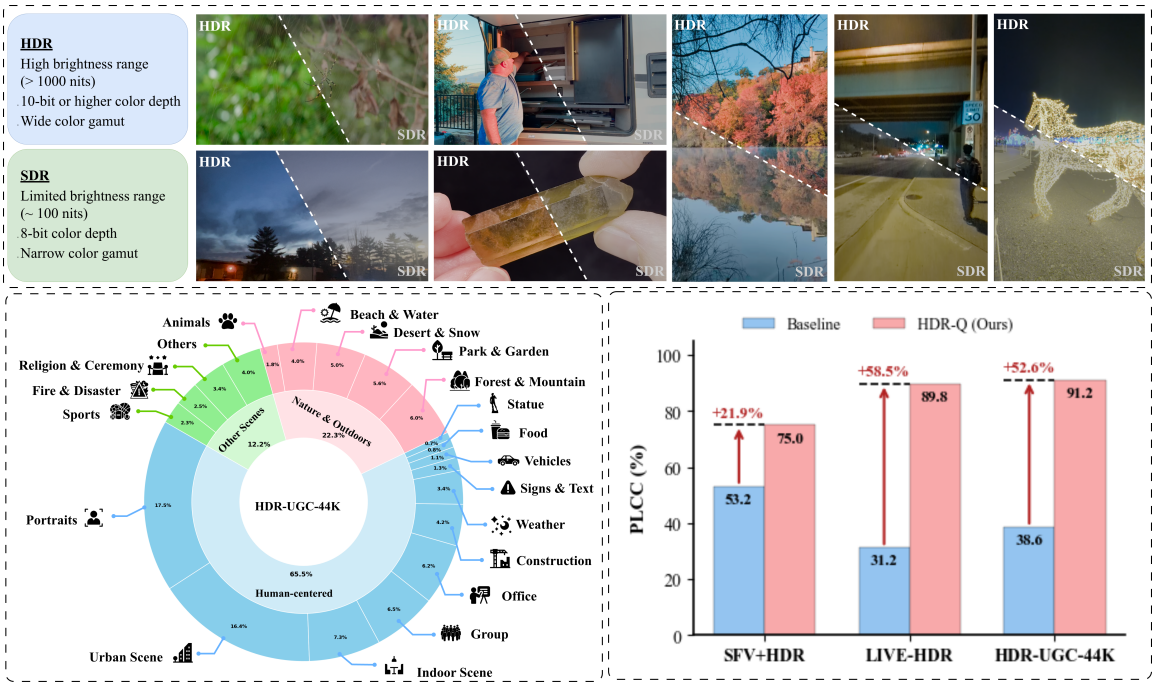


Figure 1. **Overview of our dataset and performance evaluation.** Top: Example comparisons between HDR and SDR frames, illustrating differences in brightness range, color depth, and visual detail across diverse scenes. Bottom-left: The distribution of video categories in the Beyond8Bits dataset, covering human-centered content, nature & outdoor scenes, and various other real-world scenarios. Bottom-right: Performance comparison between our proposed HDR-Q model and baseline methods on three datasets, where HDR-Q achieves significant improvements in PLCC.

Abstract

High Dynamic Range (HDR) user-generated (UGC) videos are rapidly proliferating across social platforms, yet most perceptual video quality assessment (VQA) systems remain tailored to Standard Dynamic Range (SDR). HDR's higher bit depth, wide color gamut, and elevated luminance range expose distortions such as near-black crushing, highlight clipping, banding, and exposure flicker that amplify UGC artifacts and challenge SDR models. To catalyze progress, we curate ***Beyond8Bits****, a large-scale subjective dataset of ∼44K videos from 6.5K sources with >1.5M crowd ratings, spanning diverse scenes, capture conditions, and compression settings. We further introduce* ***HDR-Q****, the first Multimodal Large Language Model (MLLM) for HDR-UGC VQA. We propose (i) a novel HDR-aware vision encoder to produce HDR-sensitive embeddings, and (ii) HDR-Aware Policy Optimization (HAPO), an RL finetuning framework that anchors reasoning to HDR cues. HAPO augments GRPO via an HDR–SDR contrastive KL that encourages token reliance on HDR inputs and a gaussian weighted regression reward for fine-grained MOS*

calibration. Across Beyond8Bits and public HDR-VQA benchmarks, HDR-Q delivers state-of-the-art performance. Project page: https://shreshthsaini.github.io/Beynod8Bits.

1. Introduction

The digital media ecosystem has been transformed by the explosive growth of user-generated content (UGC) on platforms such as YouTube, TikTok, and Instagram [1, 39, 42]. In parallel, High Dynamic Range (HDR) video has become mainstream, offering higher bit depth, wider color gamut, and extended luminance range compared to Standard Dynamic Range (SDR) [8, 49, 50, 53]. These characteristics enhance perceptual realism but also accentuate distortions that are less visible in SDR such as near-black crushing, highlight clipping, banding, and exposure flicker often compounded by compression or capture artifacts common in UGC (Fig. 2). As a result, evaluating the perceptual quality of HDR-UGC content remains an open and underexplored challenge. Existing Video Quality Assessment (VQA) models struggle in this regime. Methods trained on professionally generated HDR datasets [8, 53] or SDR-UGC videos [31, 76] fail to generalize to the heterogeneous capture conditions, device variations, and uncontrolled distortions of real-world HDR-UGC. Moreover, current HDR subjective datasets are small and limited in scope, focusing primarily on synthetic distortions or curated professional content [17, 58]. This lack of large-scale, real-world HDR-UGC annotations represents a major obstacle to developing models that align with human perceptual judgments.

Meanwhile, multimodal large language models (MLLMs) have emerged as powerful reasoning systems that unify perception and language, showing promise for explainable image and video quality assessment [27, 66–69, 72, 73]. However, their direct application to HDR-UGC VQA faces key obstacles: (i) standard visual encoders are pre-trained on SDR data and fail to capture HDR-specific cues; (ii) obtaining accurate, continuous MOS predictions remains challenging within the next-token prediction paradigm, as discrete-level or regression-head approaches [14, 66, 69, 73, 80] lack fine-grained calibration; (iii) without explicit incentives, policies often neglect HDR inputs and rely on textual priors, a form of modality neglect [10, 59, 78].

To address these challenges, we introduce Beyond8Bits, the first large-scale, crowdsourced subjective HDR-UGC quality dataset containing ∼44K videos from 6,861 diverse sources with over 1.5M human ratings. This dataset provides the necessary foundation for training and evaluating models that reflect real-world HDR perceptual phenomena. Building upon it, we propose **HDR-Q**, the first multimodal large language model specifically designed for HDR-UGC quality assessment. HDR-Q integrates two novel components: (i) an HDR-aware vision encoder that learns HDR-sensitive representations while maintaining semantic alignment, and (ii) HDR-Aware Policy Optimization (HAPO), a reinforcement learning framework that enforces HDR grounding through an HDR–SDR contrastive KL term, stabilizes entropy, and refines token-level credit assignment via entropy-weighted advantages. A Gaussian regression reward further enables fine-grained MOS calibration, while group-level self-rewarding improves reasoning consistency. Our contributions are:

- We introduce Beyond8Bits, the largest subjective HDR-UGC quality dataset.
- We propose **HDR-Q**, a novel MLLM-based VQA model that combines the new HDR-aware vision encoder with our novel HAPO policy, an RL finetuning paradigm tailored for perceptual reasoning under HDR conditions.
- Extensive experiments on Beyond8Bits and public HDR benchmarks demonstrate that HDR-Q achieves state-of-the-art MOS prediction and generates concise, HDR-grounded rationales, establishing a new direction for HDR-aware VQA research.

2. Related Work

2.1. HDR-VQA: Datasets & Models

Early VQA datasets such as CVD2014 [41], LIVE-VQA [52], LIVE-VQC [55], LSVQ [70], MDVQA [76], and Maxwell [64] enabled both handcrafted [15, 25, 36–38, 47] and deep VQA models [20, 26, 33, 60–62], but remain SDR-oriented and unsuitable for HDR due to fundamental differences in luminance and tone-mapping. HDR-specific subjective datasets [4, 5, 7, 43, 46] addressed this gap, though many are outdated or restricted. Recent releases such as LIVE-HDR [53] (310 annotated videos) and SFV+HDR [58] (2,000 clips, 300 rated) provide more reliable benchmarks for modern HDR algorithms. Correspondingly, HDR-VQA models have emerged: full-reference metrics HDR-VQM [40], HDR-BVQM [2], and PU21 [35] use brightness-aware or perceptually uniform transforms but rely on references and struggle with diverse HDR distortions. Blind methods such as HDR-ChipQA [18] and HIDRO-VQA [49] extend ChipQA and CONTRIQUE [33] through nonlinear luminance mappings or large-scale unlabeled HDR data. Nonetheless, existing datasets and models still fail to capture the heterogeneous degradations of HDR UGC, motivating new data and modeling strategies.

2.2. MLLM-Based Perceptual Quality Assessment

MLLMs have recently been explored for IQA/VQA. Benchmarks such as Q-Bench [65] revealed large gaps between MLLMs and human judgments, spurring instruction tuning (Q-Instruct [67]) and descriptive distortion reasoning (DepictQA, DepictQA-Wild [71, 72]). Compar-

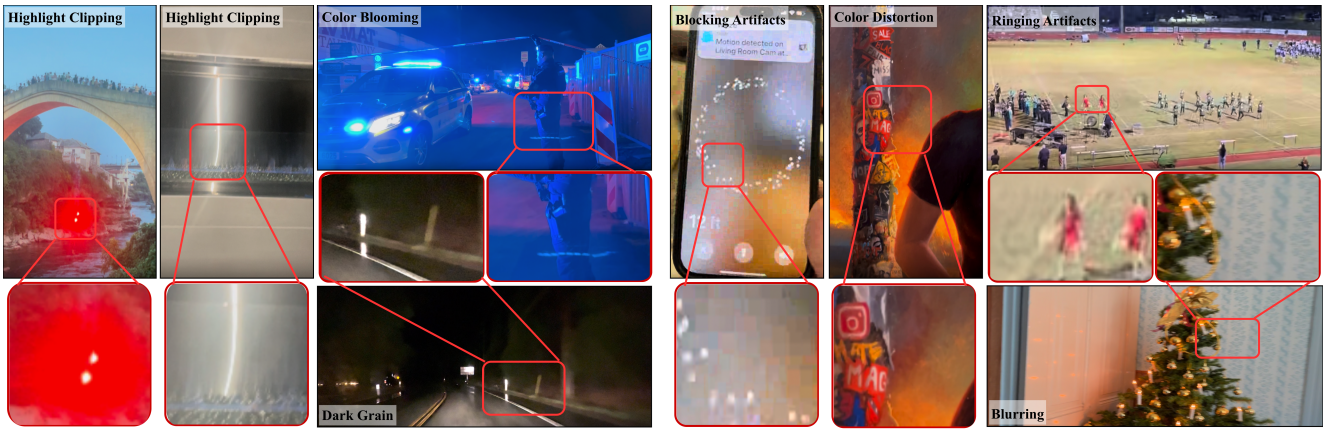


Figure 2. Typical challenges in HDR-UGC videos, including HDR-specific issues (e.g., highlight clipping, color blooming, dark grain) and UGC/compression-related artifacts (e.g., color distortion, blocking, ringing, blurring).

ative and ranking-based methods (Compare2Score [80], VisualQuality-R1 [69]) further improved human alignment, while Q-Align [66], DeQA-Score [73], and Q-Insight [27] targeted interpretability, regression fidelity, and joint degradation reasoning. Video extensions include Q-Bench-Video [77] and MVQA-68K [44], which provide large-scale, multi-dimensional annotations and textual rationales for training video-aware MLLM quality evaluators.

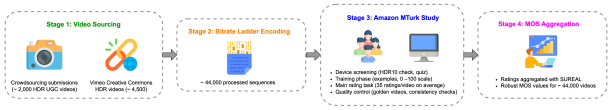


Figure 3. Pipeline of Beyond8Bits construction.

3. Dataset: Beyond8Bits

Existing HDR VQA datasets [9, 17, 50, 53, 58] are limited in scale, diversity, or dynamic range, and primarily focus on professionally produced content. In contrast, real-world HDR user-generated videos (HDR-UGC) exhibit a far wider range of luminance, motion, and compression characteristics, often captured under uncontrolled conditions. To bridge this gap, we introduce **Beyond8Bits**, the largest and most diverse HDR VQA dataset to date, explicitly designed for real-world HDR-UGC quality assessment.

3.1. Data Collection and Processing

We collected 6,861 unique HDR source videos from two complementary sources: (1) a dedicated crowdsourcing campaign where users contributed HDR clips captured on consumer devices (iPhone, Pixel, Galaxy, *etc.*) under research consents, contributing 2,253 videos, and (2) public HDR videos from Vimeo licensed under Creative Commons, contributing 4,608 videos. This combination ensures coverage across human-centric, natural, and low-light scenes with rich intra and inter-device variability.

Each source video was verified for HDR metadata (PQ transfer, 10-bit HEVC, BT.2020 gamut) and filtered to remove duplicates, static frames, and unsuitable content. Clips were trimmed to a maximum of 10 seconds and transcoded under a bitrate ladder simulating real-world streaming conditions [3, 19] at multiple resolutions (1080p–360p) and bitrates (0.2–5 Mbps), see Appendix D. All versions retained full HDR signaling, producing a total of ∼44,276 processed video clips, an equal mix of landscape and portrait orientations was maintained where possible. Fig. 1 visualizes the dataset composition and example content diversity.

3.2. Subjective Quality Study

We conducted a large-scale subjective study on Amazon Mechanical Turk (AMT), marking the first HDR large-scale crowdsourced VQA study at this scale. To ensure display fidelity, only workers with verified HDR-capable devices and browsers were admitted. Our Human Intelligence Task (HIT) design incorporated several quality control measures. Each HIT began with instructions and a qualification quiz checking for HDR display capability, and understanding of the task. Participants first completed a training and calibration phase with representative examples and then rated batches of clips using a continuous 0–100 likert-scale following ITU-R BT.500-14 [23] guidelines. To ensure reliability, we embedded hidden quality control videos (repeats and golden-set videos with known quality ranges established in pilot studies). Strict participant screening and rejection criteria were applied based on consistency checks on control videos, display bit depth, internet speed, task completion times, and reported viewing conditions. Golden-

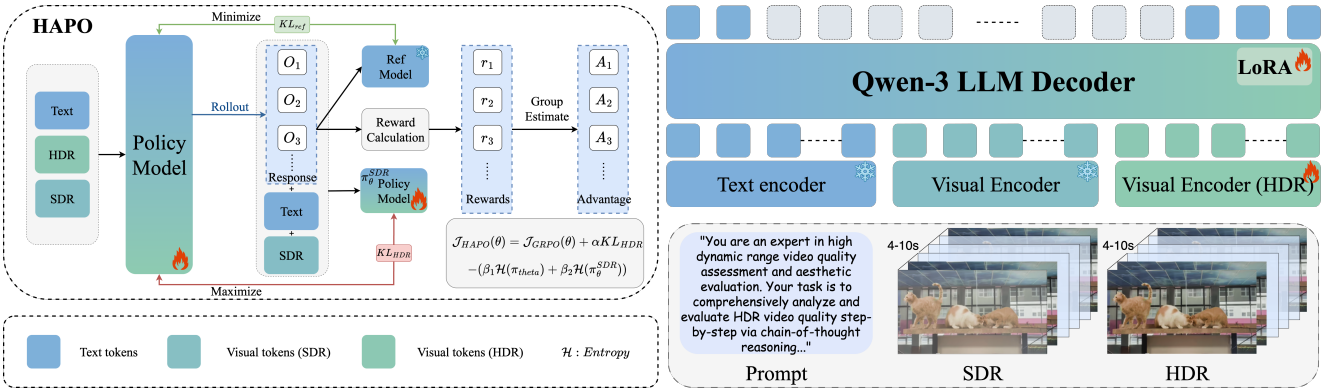


Figure 4. **Overview of HDR-Q with HAPO.** Left: HAPO compares rollouts under HDR inputs (text + SDR + HDR tokens) versus an HDR-deprived pathway (text + SDR only), maximizing their KL divergence to enforce HDR grounding and applying dual-entropy regularization to prevent reward hacking. Group-wise rewards include MOS/attribute accuracy, reasoning quality, and self-rewarding. Right: a LoRA-tuned LLM decodes the HDR-aware reasoning; visual inputs originate from both a standard encoder and our HDR-aware adapter.

set and repeat videos were embedded to assess intra and inter-subject consistency. Over 1.5M valid ratings were collected after rigorous quality control. Each video received on average ∼35 independent ratings. This large-scale design captures genuine perceptual variability under realistic HDR viewing conditions.

3.3. MOS Aggregation

To aggregate the subjective ratings into reliable MOS, we employed the Subjective Reliability (SUREAL) method [28]. SUREAL provides a Maximum Likelihood Estimate (MLE) of the true video quality (ψ_j) by modeling individual subject ratings (S_{ij}) while accounting for subject bias (Δ_i) and inconsistency (ν_i). The model is given by:

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

Parameters were estimated to maximize the log-likelihood. The resulting MOS values exhibit strong inter-subject correlation (median SRCC 0.90), confirming study reliability. Beyond8Bits spans diverse content categories (human, indoor, outdoor, night, motion-intensive), varying brightness distributions, and wide MOS coverage (10–95). Key statistics, including spatial/temporal complexity and comparisons with existing datasets, is provided in the Appendix D. Beyond8Bits provides an essential foundation for modern HDR-aware training perceptual models such as **HDR-Q**.

4. Preliminaries

Large-scale multimodal reinforcement learning requires stable optimization without the high variance of critic-based methods such as PPO [51]. Group Relative Policy Optimization (GRPO) [54] achieves this by normalizing rewards within a sampled response group, eliminating the need for a learned value network while preserving sample efficiency. GRPO [54] has become a key component of modern LLM and MLLM post-training pipelines [11, 74], particularly when direct reward modeling is infeasible.

Formulation. Given a multimodal dataset $\mathcal{D} = \{(q, I, a)\}$ with input prompt q, multimodal input I, and target answer a, we sample K candidate completions $\{o_i\}_{i=1}^K$ from the previous policy $\pi_{\theta_{\text{old}}}$. Each completion receives a scalar reward R_i, and its normalized group-relative advantage is computed as:

$$\hat{A}_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \quad \mu_R = \frac{1}{K}\sum_{j=1}^{K} R_j, \\ \sigma_R = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(R_j - \mu_R)^2}. \tag{2}$$

The clipped surrogate objective becomes:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,I)\sim\mathcal{D},\, o_i\sim\pi_{\theta_{\text{old}}}} \frac{1}{K}\sum_{i=1}^{K}\frac{1}{|o_i|}\sum_{t} \Big[\min\big(\rho_{i,t}\hat{A}_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\big) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Big]. \tag{3}$$

where $\rho_{i,t} = \pi_\theta(o_{i,t}|q, I, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t}|q, I, o_{i,<t})$ denotes the token-level importance ratio, and π_{ref} is a frozen reference policy that anchors stability.

Limitations. Despite its stability, vanilla GRPO [54] lacks explicit mechanisms to ensure that the learned policy grounds its behavior in perceptual cues from the input modality [59]. In perception-heavy tasks such as HDR-UGC VQA, this leads to modality neglect [59, 78], where the policy achieves high textual coherence yet ignores HDR

visual information. It also treats all output tokens equally, disregarding token-level uncertainty and reasoning structure issues critical in multimodal reasoning tasks. These limitations motivate our proposed HDR-Aware Policy Optimization (HAPO) (Sec. 5.2), which extends GRPO [54] with HDR–SDR contrastive grounding, dual-entropy regularization, and entropy-weighted advantage shaping.

5. Method: HDR-Q

We introduce **HDR-Q**, a multimodal large language model (MLLM) designed for perceptual quality assessment of HDR user-generated videos. The framework couples an HDR-aware vision encoder with a reinforcement learning (RL) objective, HDR-Aware Policy Optimization (HAPO), that explicitly enforces HDR grounding, stabilizes learning against reward hacking, and improves reasoning fidelity. As illustrated in Fig. 4, HDR-Q integrates both perceptual and reasoning pathways: (i) the HDR-aware encoder yields HDR-sensitive embeddings that capture luminance extremes and color-volume fidelity, while (ii) HAPO finetunes the policy to rely on these cues through contrastive, entropy-regularized RL.

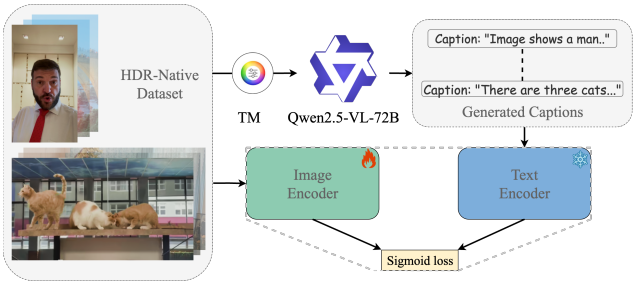


Figure 5. **HDR-aware vision encoder finetuning.** We adapt SigLIP-2 [57] using HDR–SDR frame–caption pairs with captions generated by Qwen2.5-VL-72B, promoting perceptually aligned HDR embeddings.

5.1. HDR-Aware Vision Encoder

Let $v = \{x_t\}_{t=1}^{T}$ denote a 10-bit HDR video in PQ (BT.2020). We preserve the HDR signal at full precision avoiding tone compression to retain near-black structure, highlight dynamics, and wide-gamut color relationships. For contrastive supervision, an SDR counterpart $v^{SDR} = \{TM(x_t)\}_{t=1}^{T}$ is obtained via a deterministic tone-mapping operator $TM(\cdot)$ (PQ$\rightarrow \gamma$ mapping, quantization, and BT.709 contraction).

We adapt a pretrained SigLIP-2 encoder [57] $\mathcal{E}_\psi$ on HDR frame–caption pairs (x_t, c_t), where captions are generated by Qwen2.5-VL-72B [6]. The goal is to yield embeddings that remain semantically aligned yet intrinsically sensitive to HDR variations.

Dual-Domain Supervision. A key challenge is that generic captions c_t are equally valid for x_t and x_t^{SDR}, potentially causing collapse where HDR and SDR embeddings overlap. To avoid this, we introduce dual-domain supervision. For each HDR frame x_t, we generate x_t^{SDR} and enforce contrastive separation: the HDR embedding must remain closer to its caption than the SDR embedding:

$$\mathcal{L}_{\text{contrast}} = \max\Big(0,\ \delta - D\big(\mathcal{E}_\psi(x_t),\ \mathcal{E}_\psi(c_t)\big) + D\big(\mathcal{E}_\psi(x_t^{SDR}),\ \mathcal{E}_\psi(c_t)\big)\Big). \tag{4}$$

where $D(\cdot,\cdot)$ denotes cosine distance and δ is a margin. The full encoder loss combines alignment and HDR discrimination:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{Sigmoid}}(x_t, c_t) + \lambda_{\text{ctr}}\mathcal{L}_{\text{contrast}} \tag{5}$$

ensuring that the learned embeddings remain semantically faithful while being perceptually attuned to HDR contrast and luminance cues (see Fig. 5).

5.2. HDR-Aware Policy Optimization (HAPO)

While GRPO [54] stabilizes multimodal RL, it offers no guarantee that the policy exploits visual cues rather than textual priors. HAPO extends GRPO [54] with three HDR-specific components that explicitly enforce modality grounding (See Appendix E):

(i) HDR–SDR Contrastive KL. To prevent modality neglect [78], we contrast rollouts with and without HDR tokens:

$$\mathcal{K}_{\text{HDR}}(\theta) = D_{\text{KL}}\big(\pi_\theta^{HDR} \,\|\, \pi_\theta^{SDR}\big) \tag{6}$$

where π_θ^{HDR} and π_θ^{SDR} are policies with and without HDR input. Maximizing $\mathcal{K}_{\text{HDR}}$ ensures that removing HDR tokens significantly perturbs the decoding distribution, thereby incentivizing the model to exploit HDR-specific information rather than collapsing into SDR-only reasoning.

(ii) Dual-Entropy Regularization. A well-known pitfall in contrastive KL maximization is entropy inflation, the policy can trivially satisfy the objective by producing overly uncertain outputs [45, 75]. To prevent this, we introduce policy entropy regularization on both HDR and SDR pathways:

$$\mathcal{H}_{\text{dual}}(\theta) = \mathbb{E}_{o\sim\pi_{\theta_{old}}} \frac{1}{K}\sum_{i,t}\Big[\eta_1\,\mathcal{H}\big(\pi_\theta^{HDR}(o_{i,t})\big) + \eta_2\,\mathcal{H}\big(\pi_\theta^{SDR}(o_{i,t})\big)\Big]. \tag{7}$$

where $\mathcal{H}$ denotes token-level entropy, i.e. $\mathcal{H}(\pi_\theta) = \log \pi_\theta$, and η_1 and η_2 are hyperparameters. This prevents collapse while preserving sharp, HDR-grounded distributions.

(iii) High-Entropy Weighting (HEW). GRPO assigns the same normalized advantage $\hat{A}_i$ to all tokens of a completion o_i, regardless of their informativeness. However, recent work [12] demonstrates that reinforcement learning benefits from focusing policy gradients of tokens promoting exploration, and thus improving reasoning, while tokens following fixed reasoning path provide little signal. In HDR-UGC VQA, high-entropy tokens typically occur when the model must identify or calibrate HDR-specific distortions (e.g., banding in gradients, highlight clipping, near-black crushing). By amplifying the learning signal at these tokens, HEW directs policy optimization toward the most informative reasoning steps, yielding stronger HDR grounding and more precise MOS predictions. We then rescale the group-normalized advantage $\hat{A}_i$ into a token-specific advantage:

$$w_{i,t} = \mathrm{clip}\left(1 + \lambda_{\mathrm{HEW}} \frac{H_{i,t}}{\frac{1}{|o_i|}\sum_{t'=1}^{|o_i|} H_{i,t'}}, w_{\min}, w_{\max}\right),$$
$$\tilde{A}_{i,t} = w_{i,t} \cdot \hat{A}_i. \tag{8}$$

where $H_{i,t}$ is per-token entropy.

Full HAPO Objective. Combining these terms yields:

$$\mathcal{J}_{\mathrm{HAPO}}(\theta) = \mathbb{E}_{o\sim\pi_{\theta_{\mathrm{old}}}}\left[\frac{1}{K}\sum_{i,t}\min\Big(\rho_{i,t}\tilde{A}_{i,t}, \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\tilde{A}_{i,t}\Big)\right] - \beta\, D_{\mathrm{KL}}\big(\pi_\theta^{HDR} \,\|\, \pi_{\mathrm{ref}}\big) + \gamma\, \mathcal{K}_{\mathrm{HDR}}(\theta) - \mathcal{H}_{\mathrm{dual}}(\theta). \tag{9}$$

This enforces HDR-aware reasoning while maintaining stable optimization.

Mutual Information Perspective. Our HDR–SDR contrastive KL can be interpreted as enforcing an information-theoretic dependency between HDR inputs and model outputs. Let v denote the HDR video, v^{SDR} its SDR tone-mapped counterpart, and o the output sequence. By applying variational mutual information bounds [24, 32], we obtain $\mathbb{E}_{v,v^{SDR}}[\mathcal{K}_{\mathrm{HDR}}(\theta)]$ as:

$$\mathbb{E}_{v,v^{SDR}}\,\mathbb{E}_{o\sim\pi_\theta(\cdot\,|\,v,v^{SDR})}\left[\log\frac{\pi_\theta(o\,|\,v,v^{SDR})}{\pi_\theta(o\,|\,v^{SDR})}\right] \geq I_\theta\big(o;\, v, v^{SDR}\,\big|\,v^{SDR}\big) - \kappa_\theta. \tag{10}$$

where $I_\theta(o; v, v^{SDR}|v^{SDR})$ is the conditional mutual information under π_θ, and κ_θ captures mismatch due to conditioning on v^{SDR}. This result shows that maximizing equation 6 provably increases HDR informativeness, ensuring the policy relies on HDR-specific cues rather than collapsing to SDR-only reasoning.

Table 1. Performance on Beyond8Bits. Best results in **blue bold**, second best are underlined

Model	SRCC(↑)	PLCC(↑)	RMSE(↓)	KRCC(↑)
DL models				
BRISQUE [36]	0.4096	0.4689	11.7019	0.2797
CONTRIQUE [33]	0.6245	0.6054	15.0224	0.4464
RE-IQA [48]	0.5698	0.5441	17.9049	0.4038
VBLIINDS [47]	0.4440	0.4397	11.7234	0.3044
CONVIQT [34]	0.7987	0.8099	8.4807	0.6095
FastVQA [60]	0.4909	0.4193	26.1325	0.3398
FasterVQA [61]	0.4808	0.3224	29.6357	0.3367
DOVER [62]	0.5094	0.5037	16.7176	0.3548
COVER [20]	0.6645	0.6645	16.8597	0.4870
HDRMAX [53]	0.6054	0.6070	10.1400	0.4277
HDRChipQA [18]	0.7180	0.7290	8.2987	0.5282
HIDROVQA [49]	0.8508	0.8784	6.0875	0.6694
MLLM base model				
Qwen2.5-VL(7B) [6]	0.3089	0.3228	27.8899	0.2432
GLM-4.1V-Thinking(9B) [21]	0.2641	0.3944	23.8883	0.2924
Ovis2.5(9B) [30]	0.3423	0.3860	26.7570	0.2823
OmniLong-Qwen2.5-VL(7B) [56]	0.3472	0.3595	25.7616	0.2677
MLLM VQA model				
Q-Align [66]	0.4615	0.3673	20.3411	0.3257
Q-Insight [27]	0.5170	0.5621	20.7832	0.4138
Q-Instruct [27]	0.5035	0.4712	19.6567	0.3496
DeQA [73]	0.5064	0.4642	19.5772	0.3586
Visual-Quality-Q1 [69]	0.3909	0.3617	23.6462	0.2809
HDR-Q (SDR)	0.8914	0.8895	7.4240	0.7052
HDR-Q (full)	**0.9206**	**0.9118**	**5.1594**	**0.7218**

5.3. Rewards and Training Pipeline

HAPO jointly optimizes three reward signals: format (R_{fmt}), regression accuracy (R_{sc}) [27, 69, 72], and self-consistency (R_{self}) [79] combined as

$$\mathcal{R}_i = w_{\mathrm{fmt}} R_{\mathrm{fmt}} + w_{\mathrm{sc}} R_{\mathrm{sc}} + w_{\mathrm{self}} R_{\mathrm{self}}. \tag{11}$$

A Gaussian-weighted score reward stabilizes fine-grained MOS prediction, while the self-reward consolidates within-group consensus.

Two-Stage RL Training. Our training follows a two-stage RL-based paradigm [10, 13], both optimized with the same objective but serving distinct purposes:

- **Stage 1 (Modality Alignment):** aligns HDR tokens and projection layers via short HAPO runs.
- **Stage 2 (Full-RFT):** applies complete HAPO optimization on the HDR-UGC corpus, balancing distortion diversity and reasoning quality.

Overall, **HDR-Q** unifies perceptual sensitivity and reasoning stability, the HDR encoder injects physical luminance awareness, contrastive KL enforces grounding, entropy regularization curbs uncertainty, and HEW refines token-level learning yielding accurate, interpretable HDR-aware quality judgments.

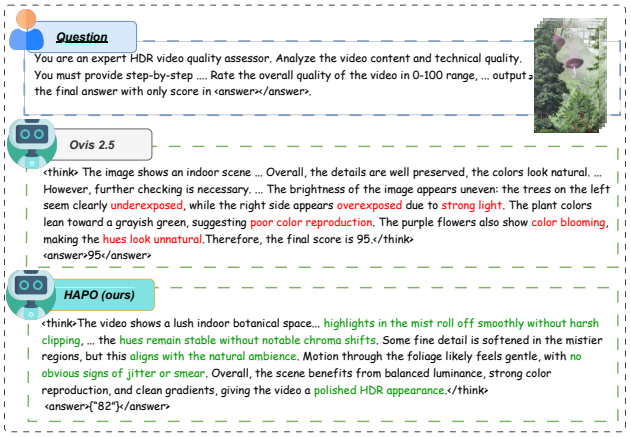


Figure 6. Given the same HDR video, OVIS 2.5 produces multiple incorrect judgments. In contrast, our HAPO-enhanced HDR-Q provides HDR-grounded reasoning. (Best viewed zoomed in)

6. Experiments

6.1. Experimental Setup

Datasets. We evaluate HDR-Q on the curated Beyond8Bits benchmark and test generalization on two public HDR-VQA datasets: LIVE-HDR [53] and SFV+HDR [58]. Beyond8Bits is split by source identity into 70%/20%/10% train/val/test to avoid overlap.

Metrics. Following VQA convention [31, 34, 49, 50], we report Spearman's Rank (SRCC), Pearson's Linear (PLCC), and Kendall's Rank (KRCC) correlations (↑ higher is better), and RMSE (↓ lower is better) against MOS.

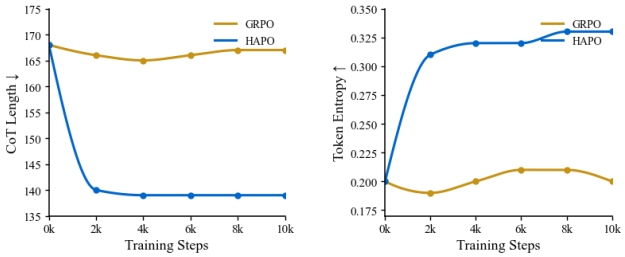
Baselines. We compare four category of methods. (i) NR-VQA: BRISQUE [36], VBLIINDS [47], FastVQA [63], FasterVQA [61], DOVER [62], CONVIQT [34], COVER [20]; (ii) HDR-VQA: HDRMAX [16], HDR-ChipQA [18], HIDRO-VQA [49]; (iii) MLLM/VLM-VQA: Q-Align [66], Q-Instruct [67], Q-Insight [27], DeQA [73], Visual-Quality-R1 [69]; (iv) Base MLLMs: Qwen2.5-VL [6], GLM-4.1V-Thinking [21], Ovis2.5 [30], OmniLong-Qwen2.5-VL [56]. Where applicable, methods are re-trained on Beyond8Bits using authors' protocols; others are evaluated in their released form.

Implementation details. HDR-Q is built on Ovis2.5 [30] with rank-4 LoRA adapters [22]. Frames are ingested at native 10-bit PQ (no linear downscaling). Each clip is uniformly sampled into $T = 8$ frames; visual tokens from $\mathcal{E}_\psi(x_t)$ and SDR tokens from $\mathcal{E}_\psi(x_t^{\text{SDR}})$ feed the language decoder via learned projections. In HAPO, group size $K = 8$; clip range $\epsilon = 0.1$ (clip-higher); reference KL weight $\beta = 0.02$; HDR–SDR contrastive KL weight $\gamma = 0.5$; policy entropy $\eta_1, \eta_2 = 0.01, 0.05$; HEW modulation $\lambda_{\text{HEW}} = 0.3$ with $w_{\min} = 0.5, w_{\max} = 2.0$. In gaussian score reward R_{sc} we use $\sigma = 3$ and $\alpha = 1$; weights $(w_{\text{fmt}}, w_{\text{sc}}, w_{\text{self}})$ tuned on validation. We use AdamW [29], lr 1×10^{-5}, batch size of 4. We use four NVIDIA H200 GPUs for training.

6.2. Main Results

Overall performance on Beyond8Bits. Table 1 reports quantitative results on Beyond8Bits. HDR-Q consistently outperforms all SDR, HDR, and MLLM-based baselines across correlation metrics, with substantial gains in RMSE. Against HDR-ChipQA [18] and HIDRO-VQA [49], HDR-Q achieves higher SRCC/PLCC with lower RMSE, indicating the benefits of HDR-aware embeddings plus HAPO grounding. Against FastVQA [63] and DOVER [62], HDR-Q remains robust despite diverse UGC capture pipelines. Relative to all MLLM and VLM models, HDR-Q's gains stem from HDR-aware encoder finetuning on 10-bit PQ without linear SDR scaling, and HDR–SDR contrastive KL (prevents modality neglect). To test robustness, we evaluate zero-shot transfer on LIVE-HDR [53] and SFV+HDR [58] (Table 2). HDR-Q retains high correlation and low error without retraining evidence that its HDR-aware encoder and HAPO grounding produce representations that generalize across UGC and PGC HDR domains.

HDR-Q generates concise, HDR-aware reasoning (Fig. 6), detecting "natural indoor scene," "possible hues from chroma shifts," or "jitter" and linking them to perceptual judgments. Fig. 7 shows that HAPO stabilizes CoT length while HEW concentrates gradients on informative tokens yielding efficient and interpretable reasoning.



((a)) CoT Length ((b)) Token Entropy

Figure 7. Analysis of Chain-of-Thought (CoT) Length and Token Entropy over training iterations. (a) shows the decrease in CoT length, while (b) shows the corresponding increase in token entropy.

6.3. Ablation Studies

Table 3 quantifies the contribution of each component. Removing HDR finetuning drops SRCC markedly, confirming that 10-bit cues are essential. Omitting HDR–SDR KL causes modality neglect, while disabling entropy regularization yields unstable, verbose reasoning. HEW improves token-level credit assignment, and self-rewarding enhances stability on noisy samples.

Table 2. Cross-dataset performance Comparison on LIVE-HDR [53] and SFV+HDR [58] Datasets.

Model	LIVE-HDR				SFV+HDR			
	SROCC(↑)	PLCC(↑)	RMSE(↓)	KRCC(↑)	SROCC(↑)	PLCC(↑)	RMSE(↓)	KRCC(↑)
DL models								
BRISQUE [36]	0.7251	0.7139	12.6404	0.3424	0.4664	0.4186	0.3811	0.3165
CONTRIQUE [33]	0.8170	0.7875	11.2514	0.5876	0.5901	0.5959	0.3368	0.4204
RE-IQA [48]	0.7196	0.6883	15.1653	0.5197	0.5822	0.5998	0.3072	0.4145
VBLIINDS [47]	0.7483	0.7193	12.7794	0.2541	0.3335	0.2713	0.3988	0.2300
CONVIQT [34]	0.7922	0.8001	11.9681	0.6041	0.5736	0.6017	0.3412	0.4170
FastVQA [60]	0.5182	0.5727	18.8379	0.3822	0.7130	0.7295	0.7467	0.5193
FasterVQA [63]	0.3385	0.4114	22.1425	0.2282	0.6948	0.6889	0.3081	0.5089
DOVER [62]	0.6303	0.6832	17.0005	0.4692	0.6001	0.6154	0.5750	0.4270
COVER [20]	0.5022	0.5013	21.3297	0.3731	0.6613	0.7048	0.6831	0.4705
HDRMAX [53]	0.6308	0.5088	15.4146	0.4509	0.5371	0.5463	0.3495	0.3821
HDRChipQA [18]	0.8250	0.8344	9.8038	0.4501	0.6296	0.6508	0.3271	0.4440
HIDROVQA [49]	0.8793	0.8678	8.8743	0.6919	0.7003	0.7320	0.2735	0.5156
MLLM base model								
Qwen2.5-VL(7B) [6]	0.3099	0.3630	30.2082	0.2411	0.2925	0.2696	0.7480	0.2270
GLM-4.1V-Thinking(9B) [21]	0.4513	0.5517	26.3800	0.3464	0.5971	0.6066	0.4591	0.4484
Ovis2.5(9B) [30]	0.2948	0.3124	29.7789	0.2154	0.5909	0.5317	0.7016	0.4528
OmniLong-Qwen2.5-VL(7B) [56]	0.2403	0.2223	29.7394	0.1853	0.2363	0.2212	0.7403	0.1823
MLLM VQA model								
Q-Align [66]	0.3346	0.3604	19.8287	0.2313	0.6968	0.6709	0.5097	0.4991
Q-Insight [27]	0.3675	0.3825	25.0578	0.2820	0.6266	0.4685	0.6636	0.4747
Q-Instruct [67]	0.4083	0.4340	23.1015	0.2839	0.5830	0.5501	1.0250	0.3975
DeQA [73]	0.3321	0.3809	19.3193	0.2298	0.6850	0.6721	0.4452	0.4845
Visual-Quality-Q1 [69]	0.4824	0.5394	20.8971	0.3564	0.5955	0.5577	0.5878	0.4416
HDR-Q (SDR)	0.8542	0.8445	12.4121	0.6681	0.6971	0.7019	0.3075	0.4885
HDR-Q (full)	**0.9081**	**0.8978**	**7.6031**	**0.7363**	**0.7251**	**0.7502**	**0.2514**	**0.5261**

Table 3. **Component ablation on Beyond8Bits.** ✓=enabled, ✗=disabled. CoT len: CoT length and Tok. H: mean token entropy.

Variant	HDR-Enc.	HAPO	HDR–SDR KL	Dual Ent.	HEW	Self-R.	PLCC	SRCC	RMSE	KRCC	CoT len	Tok. H
GRPO baseline	✗	✗	✗	✗	✗	✗	0.79	0.81	10.73	0.56	168	0.20
GRPO + HDR-Enc.	✓	✗	✗	✗	✗	✗	0.81	0.83	8.96	0.61	161	0.24
HAPO w/o HDR–SDR KL	✓	✓	✗	✓	✓	✓	0.84	0.86	7.10	0.64	142	0.29
HAPO w/o Dual Ent.	✓	✓	✓	✗	✓	✓	0.89	0.91	5.82	0.71	148	0.26
HAPO w/o HEW	✓	✓	✓	✓	✗	✓	0.87	0.88	6.11	0.68	155	0.27
HAPO w/o Self-Reward	✓	✓	✓	✓	✓	✗	0.90	0.92	5.22	0.71	140	0.31
HDR-Q (Full)	✓	✓	✓	✓	✓	✓	**0.91**	**0.92**	**5.15**	**0.72**	**137**	**0.33**

Reasoning dynamics. Fig. 7 shows that HAPO reduces unnecessary CoT length over time while maintaining or improving accuracy, suggesting better use of visual evidence rather than increase in boilerplate rationales.

6.4. Complexity and Throughput

HAPO only adds an additional SDR-path forward pass only during training. Inference cost equals a single HDR path decode, maintaining competitive throughput on NVIDIA H200 GPUs.

7. Conclusion

We tackled the critical challenge of perceptual quality assessment for the fast-growing domain of HDR-UGC videos. We introduced Beyond8Bits, the largest crowdsourced subjective dataset for real-world HDR content, spanning diverse scenes, devices, and compression settings. We further proposed HDR-Q, the first MLLM for HDR VQA, combining our novel HDR-aware vision encoder with HDR-Aware Policy Optimization (HAPO), a reinforcement learning framework that enforces HDR–SDR perceptual grounding and stabilizes reasoning via dual-entropy regularization and entropy-weighted credit assignment. HAPO enables accurate, interpretable, and HDR-sensitive quality reasoning. HDR-Q achieves state-of-the-art alignment with human opinion scores across Beyond8Bits, LIVE-HDR, and SFV+HDR. By releasing the dataset, we hope to catalyze future research in HDR-aware perception, evaluation, and generative model alignment.

8. Acknowledgment

This work was supported by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844. The authors thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing VISTA compute infrastruc-

ture that contributed to the part of research outcomes in this paper.

References

[1] 99Firms. Facebook video statistics, 2024. [Online]. 2

[2] Naima Aamir, Junaid Mir, Imran Fareed Nizami, Furqan Shaukat, and Muhammad Majid. Hdr-bvqm: High dynamic range blind video quality model. *Multimedia Tools and Applications*, 80:27701 – 27715, 2021. 2

[3] Apple Inc. Hls authoring specification for apple devices, 2024. Accessed: Feb. 2024. 3

[4] Shahrukh Athar, Thilan Costa, Kai Zeng, and Zhou Wang. Perceptual quality assessment of UHD-HDR-WCG videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1740–1744. IEEE, 2019. 2

[5] Maryam Azimi et al. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 2

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 7, 8

[7] V Baroncini, K Andersson, AK Ramasubramonian, and G Sullivan. Verification test report for HDR/WCG video coding using HEVC main 10 profile. In *Proc. JCTVC-X1018 24th JCT-VC Meeting*, pages 293–303, 2016. 2

[8] Bowen Chen, Cheng-han Lee, Yixu Chen, Zaixi Shang, Hai Wei, and Alan C Bovik. Hdrsdr-vqa: A subjective video quality dataset for hdr and sdr comparative evaluation. *arXiv preprint arXiv:2505.21831*, 2025. 2

[9] Bowen Chen, Cheng-han Lee, Yixu Chen, Zaixi Shang, Hai Wei, and Alan C Bovik. Hdrsdr-vqa: A subjective video quality dataset for hdr and sdr comparative evaluation. *arXiv preprint arXiv:2505.21831*, 2025. 3

[10] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025. 2, 6

[11] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025. 4

[12] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *URL https://arxiv.org/abs/2505.22617*, 2025. 6

[13] Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *arXiv preprint arXiv:2506.00711*, 2025. 6

[14] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3206–3217, 2025. 2

[15] Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE Transactions on Image Processing*, 30:8059–8074, 2021. 2

[16] Joshua P Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C Bovik. Making video quality assessment models robust to bit depth. *IEEE Signal Processing Letters*, 30:488–492, 2023. 7

[17] 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. *IEEE Transactions on Image Processing*, 2024. 2, 3

[18] 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. *Signal Processing: Image Communication*, 129:117191, 2024. 2, 6, 7, 8

[19] Google Support. Recommended upload encoding settings, 2024. Accessed: Feb. 2024. 3

[20] Chenlong He, Qi Zheng, Ruoxi Zhu, Xiaoyang Zeng, Yibo Fan, and Zhengzhong Tu. Cover: A comprehensive video quality evaluator. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5799–5809, 2024. 2, 6, 7, 8

[21] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 6, 7, 8

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7

[23] International Telecommunication Union. Methodology for the Subjective Assessment of the Quality of Television Pictures. Technical Report BT.500-14, International Telecommunication Union, 2019. 3

[24] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv e-prints*, pages arXiv–1801, 2018. 6

[25] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019. 2

[26] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM Multimedia*, pages 2351–2359. ACM, 2019. 2

[27] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025. 2, 3, 6, 7, 8

[28] 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. 4

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[30] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 6, 7, 8

[31] 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*, pages 25963–25973, 2024. 2, 7

[32] Xiao Ma, Bingyi Kang, Zhongwen Xu, Min Lin, and Shuicheng Yan. Mutual information regularized offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:19058–19072, 2023. 6

[33] 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. 2, 6, 8

[34] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Conviqt: Contrastive video quality estimator. *IEEE Transactions on Image Processing*, 32:5138–5152, 2023. 6, 7, 8

[35] RK Mantiuk and M Azimi. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. 2021. 2

[36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 2, 6, 7, 8

[37] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[38] Anish Mittal, Michele A. Saad, and Alan C. Bovik. A completely blind video integrity oracle. *IEEE Trans. Image Process.*, 25(1):289–300, 2016. 2

[39] Maryam Mohsin. 10 youtube statistics every marketer should know in 2020, 2020. [Online]. 2

[40] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Hdr-vqm: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015. 2

[41] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. CVD2014 - A database for evaluating no-reference video quality assessment algorithms. *IEEE Trans. Image Process.*, 25(7):3073–3086, 2016. 2

[42] Omnicore. Tiktok by the numbers, 2024. [Online]. 2

[43] Xiaofei Pan, Jiaqi Zhang, Shanshe Wang, Shiqi Wang, Yun Zhou, Wenhua Ding, and Yahui Yang. Hdr video quality assessment: Perceptual evaluation of compressed hdr video. *Journal of Visual Communication and Image Representation*, 57:76–83, 2018. 2

[44] Yanyun Pu, Kehan Li, Zeyi Huang, Zhijie Zhong, and Kaixiang Yang. Mvqa-68k: A multi-dimensional and causally-annotated dataset with quality interpretability for video assessment. *arXiv preprint arXiv:2509.11589*, 2025. 3

[45] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 5

[46] Martin Rerabek, Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi. Subjective and objective evaluation of hdr video compression. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015. 2

[47] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Trans. Image Process.*, 23(3):1352–1365, 2014. 2, 6, 7, 8

[48] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5846–5855. IEEE, 2023. 6, 8

[49] 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*, pages 469–479, 2024. 2, 6, 7, 8

[50] Shreshth Saini, Alan C Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. Chug: Crowdsourced user-generated hdr video quality dataset. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2504–2509. IEEE, 2025. 2, 3, 7

[51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4

[52] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010. 2

[53] 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. 2, 3, 6, 7, 8

[54] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 4, 5

[55] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.*, 28(2): 612–627, 2019. 2

[56] Yin Song and Chen Wu. aws-prototyping/omnilong-qwen2.5-vl-7b, 2025. 6, 7, 8

[57] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil

Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5

[58] Yilin Wang, Joong Gon Yim, Neil Birkbeck, and Balu Adsumilli. Youtube sfv+ hdr quality dataset. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 96–102. IEEE, 2024. 2, 3, 7, 8

[59] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025. 2, 4

[60] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, page 538–554, Berlin, Heidelberg, 2022. Springer-Verlag. 2, 6, 8

[61] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment, 2022. 6, 7

[62] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Disentangling aesthetic and technical effects for video quality assessment of user generated content. *CoRR*, abs/2211.04894, 2022. 2, 6, 7, 8

[63] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15185–15202, 2023. 7, 8

[64] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1045–1054. ACM, 2023. 2

[65] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 2

[66] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 2, 3, 6, 7, 8

[67] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500, 2024. 2, 7, 8

[68] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer, 2024.

[69] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025. 2, 3, 6, 7, 8

[70] 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*, pages 14019–14029, 2021. 2

[71] Zhiyuan You, Jinjin Gu, Zheyuan Li, Xin Cai, Kaiwen Zhu, Chao Dong, and Tianfan Xue. Descriptive image quality assessment in the wild. *arXiv preprint arXiv:2405.18842*, 2024. 2

[72] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pages 259–276. Springer, 2024. 2, 6

[73] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 2025. 2, 3, 6, 7, 8

[74] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 4

[75] Qingcheng Zeng, Mingyu Jin, Qinkai Yu, Zhenting Wang, Wenyue Hua, Zihao Zhou, Guangyan Sun, Yanda Meng, Shiqing Ma, Qifan Wang, et al. Uncertainty is fragile: Manipulating uncertainty in large language models. *arXiv preprint arXiv:2407.11282*, 2024. 5

[76] 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. 2

[77] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmark the video quality understanding of lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3229–3239, 2025. 3

[78] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025. 2, 4, 5

[79] Xin Zhou, Yiwen Guo, Ruotian Ma, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-consistency of the internal reward models improves self-rewarding language models. *arXiv preprint arXiv:2502.08922*, 2025. 6

[80] Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai,

Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629, 2024. 2, 3