HDR-VLM: HDR-Domain Adaptation of VLMs and Preference-Aligned Quality Assessment for HDR Video Color Grading

Supplementary Material

8. More Details of Our Subjective Dataset

This section expands on the dataset details mentioned in Sec. 3.2. The collection of SDR videos is derived from 50 high-quality source clips covering a wide range of genres, including films, TV series, documentaries, and animations. These clips feature a balanced mix of indoor and outdoor environments with varying luminance conditions. In terms of technical specifications, all sources are unified to a resolution of 4K and processed into HDR formats using PQ HDR 10 [21]. The final dataset consists of 700 videos, including the 50 SDR references and 650 HDR versions generated through various grading methods.

To capture the complex perceptual variations inherent in real-world production, distinct HDR color grading schemes are employed, broadly categorized into two primary classes. *Algorithmic grading* focuses on computational conversion and technical attributes, encompassing automated sequential pipelines, deep inverse tone mapping [5] approaches, and systematic parameter adjustments such as fixed tone mapping. *Professional manual grading* is conducted by experienced artists to reflect specific artistic intents, involving the emulation of streaming platform styles, manual manipulation of luminance curves, and fine-grained tuning of color and contrast attributes.

9. More Details of Our Prompt

As mentioned in Sec. 6.2, all VLM-based systems use a common instruction template, and for each HDR-SRC pair the following fixed prompt is adopted, with only the parameter-specific phrases changed: *"Please analyze strictly according to the following format, combining images and technical parameters: Place the analytical reasoning process within <think></think>(be sure to combine image content and parameter analysis), and directly return the relative quality score of image B relative to image A within <answer></answer>. As an HDR color grading expert, please return the relative quality score of B relative to A based on visual impression, content, and relevant parameters (please only return floating-point scores within the range of [-6, 10]; please return appropriate scores, negative scores represent the inferiority of image B compared to A, and positive scores represent the superiority of B is compared to A). A parameters: (highlight area ratio = 3.3%, average brightness (nits) = 42.3, midtone contrast = 10.9, average saturation (Chroma) = 16.8, peak brightness (nits) = 712.5, shadow detail retention (shadow variance) = 0.01); B parameters: (highlight area ratio = 0.2%, average brightness (nits) = 15.4, midtone contrast = 3.7, average saturation (Chroma) = 10.9, peak brightness (nits) = 490.8, shadow detail retention (shadow variance) = 0.00)."*

10. More Visualizations

Complementing the qualitative analysis in Sec. 6.5, additional case studies are presented to illustrate the progressive refinement of HDR-VLM across training stages. As shown in Fig. 6, a specific example illustrates the clear evolution of the model from the baseline to the final version. Notably, the baseline Qwen2.5-VL [3] demonstrates limited sensitivity to HDR-specific attributes, producing nearly identical evaluations for variants B and C despite the significant disparity in human preference. The Stage 1 domain-adapted model already uses the technical descriptors in Table 4 and comments on brightness, contrast, and highlight area. After Stage 2 GRPO-based [15] alignment, HDR-VLM produces step-by-step reasoning that weighs highlight gains against shadow loss and assigns a small positive score to B and a much higher score to C, consistent with human judgments. The second row in Fig. 6 presents another SRC where only the final HDR-VLM model is shown. It again yields detailed technical analysis and calibrated scores, correctly assigning a high grade to B and a negative grade to artifact-degraded C. These examples together show that Stage 1 mainly transfers HDR knowledge into the VLM, while Stage 2 aligns the subjective scoring scale with human preference across diverse scenes. The example HDR videos are named according to their ground-truth scores, and the source SDR clips are labeled SRC1 and SRC2 in order of appearance. All files are included in the same directory as this supplementary document.

11. More Details of Our Method

Dynamic K-means grading. As discussed in Sec. 5, a distribution-adaptive 1D K-means (*dynamic K-means*) is applied to cluster relative scores within each SRC group. To extract a high-confidence hierarchical structure, an automatic grading procedure is adopted. First, the data are partitioned into positive ($\Delta q > 0$) and negative ($\Delta q < 0$) groups, using the SDR score as the reference boundary. K-means clustering is then performed on each group independently, where the optimal K^* is determined via the silhouette coefficient:

$$K^* = \arg\max_K \frac{1}{N}\sum_{i=1}^{N}\frac{b_i - a_i}{\max(a_i, b_i)}, \tag{14}$$

where a_i denotes the average intra-cluster distance and b_i represents the average nearest-cluster distance. Using a dynamic threshold $\tau_{\text{clu}} = 0.5\sigma(\Delta q)$, the procedure iteratively merges singleton grades into their nearest clusters and fuses

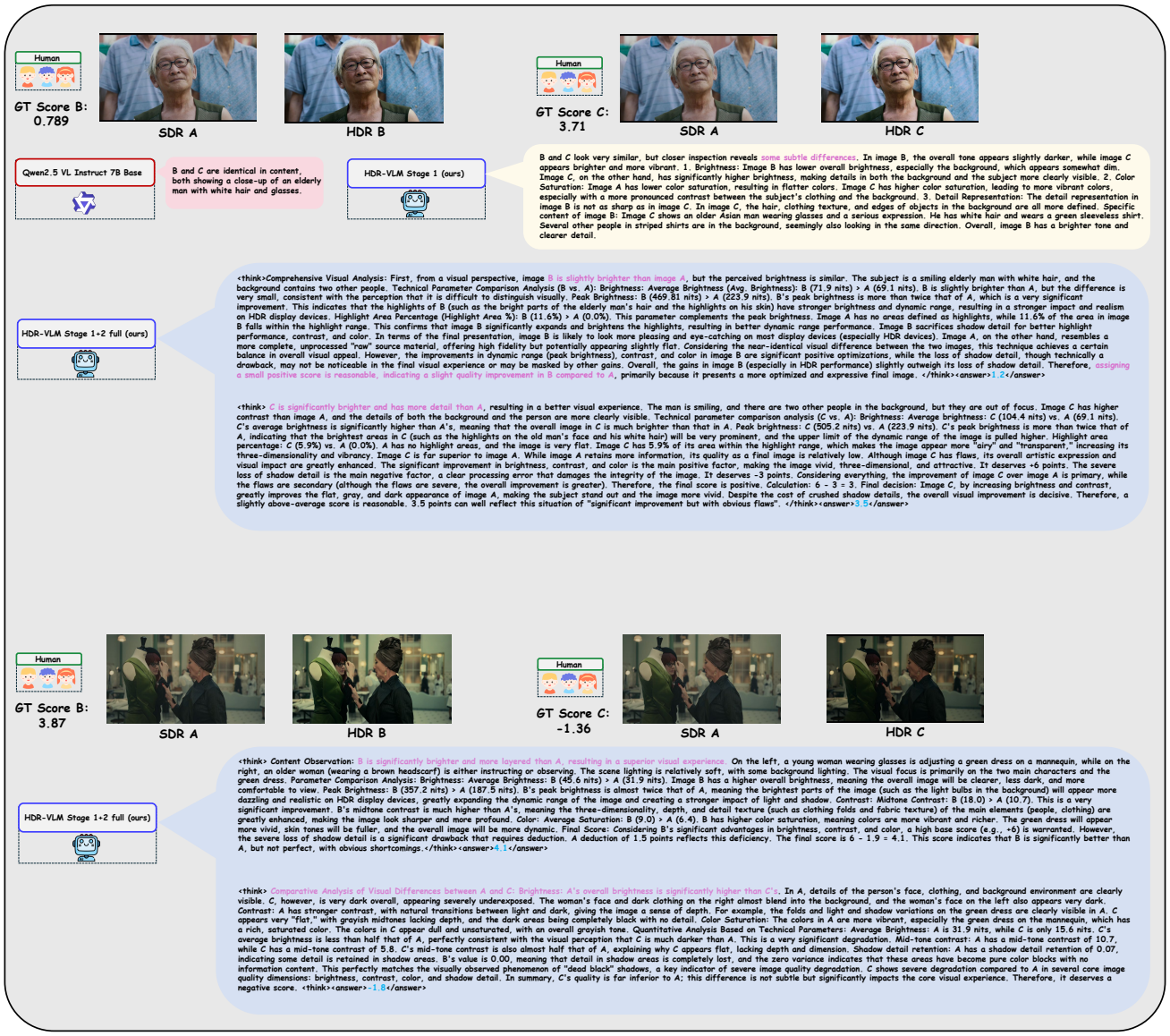

Figure 6. Qualitative case studies of the HDR-VLM pipeline from the base VLM to Stage 1 domain adaptation and Stage 2 RL alignment. The task is to score HDR color grading quality for two variants B and C relative to an SDR reference A from the same scene. In the top row, human ground-truth scores are 0.789 for B and 3.71 for C. The base Qwen2.5 VL Instruct 7B [3] model produces short generic descriptions and treats B and C as nearly identical. After Stage 1 HDR domain adaptation, the model starts to reason about brightness, contrast, and highlight area. Our final HDR-VLM model gives detailed technical analysis and assigns scores that clearly distinguish these two levels of improvement and align better with human preference. The bottom row shows an additional SRC where our final HDR-VLM again produces fine-grained technical reasoning and human-aligned scores, assigning a high score to B and a negative score to C, which is strongly degraded by artifacts. *Note: Images are tone-mapped for display and serve illustrative purposes only, not a faithful reproduction of HDR.*

adjacent clusters across the SDR boundary until convergence. Ultimately, each group yields 4–6 distinct quality grades with substantial inter-level score separations ($\Delta q \geq 2\tau_{\text{clu}}$), providing robust ranking supervision for subsequent SROCC-based rewards.

HDR descriptor tool. As mentioned in Sec. 4 and Sec. 6.2, HDR-specific technical parameters are explicitly provided to the VLM during both domain adaptation and RL alignment. These parameters are computed by an in-house analysis tool that extracts several frame-level HDR descriptors for each video frame, as summarized in Table 4. The tool computes average and peak luminance, shadow variance, mid-tone contrast, average saturation in CIELAB space, and the highlight area ratio. These physically grounded parameters jointly characterize the scene luminance level, highlight intensity, preservation of shadow detail, mid-tone contrast strength,

Table 4. Frame-level HDR descriptors used by our in-house analysis tool. All quantities are computed from decoded luminance values L_i (nits) and cover brightness, contrast and color dimensions that are strongly coupled with HDR color grading decisions.

Parameter	Formula / calculation method	Role
Average luminance	$\bar{L} = \frac{1}{N}\sum_{i=1}^{N} L_i$, where L_i is the luminance (in nits) of pixel i and N is the total number of pixels.	Reflects the average luminance level of the scene.
Peak luminance	$L_{\max} = \max(L_1, L_2, \ldots, L_N)$.	Defines the maximum luminance that the brightest point in the frame (e.g., sun, lamps, specular highlights) can reach, which is directly related to the perceived dynamic range and realism of the image.
Shadow variance	$\mathrm{Var}(L_i)$, where $L_i < P_5(L)$ and $P_p(L)$ denotes the p-th percentile of the luminance distribution L.	Evaluates shadow details. A large variance indicates rich layers and details in dark regions; a small variance means the shadows collapse into "crushed blacks" areas with severe loss of detail.
Mid-tone contrast	$\sigma(L_i)$, where $P_{30}(L) < L_i < P_{70}(L)$.	Characterizes the perceived "depth" and "clarity" of the main subjects. This quantity measures the contrast of mid-tone pixels in the primary content regions, directly affecting perceived sharpness and separation.
Average saturation	$\bar{C}^* = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(a_i^*)^2 + (b_i^*)^2}$, computed in CIELAB color space.	Captures the first color impression of the image (vivid, flat, or natural); both over-saturation and under-saturation can negatively affect perceived quality.
Highlight area ratio	$\frac{\mathrm{Count}(L_i \geq T_{\text{nits}})}{N_{\text{pixels}}} \times 100\%$, where T_{nits} is a luminance threshold in nits (225 nits in our experiments).	Describes the fraction of the image occupied by bright highlight regions and links them to the effective dynamic range of the graded video.

and overall color saturation. They are used to build the structured comparison prompts and technical parameters in our SFT and RL stages, providing interpretable factors that are closely tied to color grading quality.

Subjective experiment protocol. Following Sec. 3.2, the study uses pairwise comparisons on 1,000-nit HDR TVs with 48 observers under standard viewing conditions [43]. To improve efficiency, we employ the Hybrid-MST [27] method for the 14 video variants in each group. The process begins with expert ratings to accelerate learning and adaptively selects the most informative pairs for comparison in each round. Final quality scores are calculated using the Thurstone model, and the evaluation stops once the results stabilize.

12. More Details of Our Ablation

Table 5. Ablation on the GRPO [15] sampling hyperparameter K in Stage 2. We report PLCC and SROCC on the HDR color grading test set. Higher values indicate better performance.

Metrics	$K = 4$	$K = 5$	$K = 6$
PLCC	0.8977	0.9033	0.8945
SROCC	0.8591	0.8667	0.8587

Ablation on the sampling hyperparameter K. As described in Sec. 5.1, GRPO [15] samples K candidate responses for each input during Stage 2 policy optimization. In the HDR setting, this corresponds to sampling K candidate scores per HDR–SRC frame pair. As shown in Table 5, the performance is relatively stable when K varies from 4 to 6, and $K = 5$ yields the best PLCC/SROCC (0.9033/0.8667). Therefore, $K = 5$ is used in all main experiments.

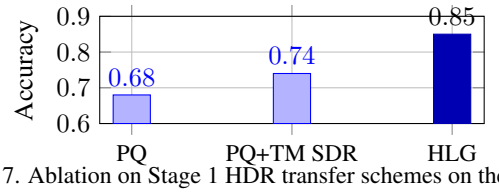

Figure 7. Ablation on Stage 1 HDR transfer schemes on the HDR domain-adaptation validation set. The task is pairwise comparison of two HDR frames from the same SRC to determine which has a larger value on a specified color grading parameter. Accuracy is reported for PQ, PQ tone-mapped SDR (PQ+TM SDR), and HLG.

Ablation on VLM adaptation. HLG [21] provides a mixed-encoding HDR domain that facilitates stable adaptation of the vision stack in VLMs, as discussed in Sec. 4. As shown in Fig. 7, an ablation compares three transfer schemes under a common training protocol that directly unfreezes and fine tunes all VLM related weights (matching the Full one-shot setting in Fig. 4), where HLG achieves a Stage 1 transfer accuracy of 0.85 compared with 0.68 for PQ [21] and 0.74 for PQ tone-mapped SDR. These results indicate that HLG serves as a more perceptually compatible bridge from SDR-centric pretraining to HDR reasoning and enables faster and more stable domain adaptation than PQ or tone-mapped SDR. Combined with the progressive schedule, Stage 1 injects HDR knowledge into the model with minimal accuracy loss on SDR content and strong training stability.