Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BrightRate-LM: Representation-Aware Reasoning Quality Assessment for User-Generated HDR Video

Shreshth Saini

saini.2@utexas.edu

Google (United States)

Yilin Wang

Google (United States)

Neil Birkbeck

Google (United States)

Balu Adsumilli

Google (United States)

Alan C. Bovik

The University of Texas at Austin

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10832291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BrightRate-LM: Representation-Aware Reasoning Quality Assessment for User-Generated HDR Video

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

[1]The University of Texas at Austin, Texas, USA.

[2]Google, Inc., Mountain View, California, USA.

Contributing authors: saini.2@utexas.edu;

Abstract

Assessing user-generated high dynamic range (HDR) video requires models to capture compression and acquisition defects across wide luminance ranges, content semantics, and time. We conduct a controlled study of multimodal large language models on the BrightVQ benchmark. Generic zero-shot models remain poorly calibrated, and supervised adaptation does not produce monotonic scaling. Multi-exposure input raises Qwen2.5-VL-3B by 0.0638 in Spearman rank-order correlation coefficient (SROCC) and lets Qwen3-VL-2B outperform Qwen3-VL-8B with Hable tone-mapped standard dynamic range input. For encoder-free Gemma-4-12B, clip-level statistics matching recovers native PQ from a defective-interface 0.0100 to 0.8383 SROCC, above the checkpoint's 0.7763 tone-mapped result, with reference-free percentile stretching reaching 0.8305. We introduce BrightRate-LM, a Qwen2.5-VL-7B model that returns an expected quality score, a visible-defect description, and reasoning that connects the evidence to the score. With multi-exposure input, it reaches 0.9052 SROCC (0.0196), compared with 0.8887 for BrightRate, and transfers without retraining at 0.8958 SROCC on the 8,281-clip official Beyond8Bits evaluation set. These results show that input representation can matter more than parameter scale and that a quality predictor can provide useful language diagnostics alongside an accurate continuous score. Project resources are available at https://shreshthsaini.github.io/BrightRate-LM/.

Keywords: HDR video, user-generated content, video quality assessment, multimodal learning

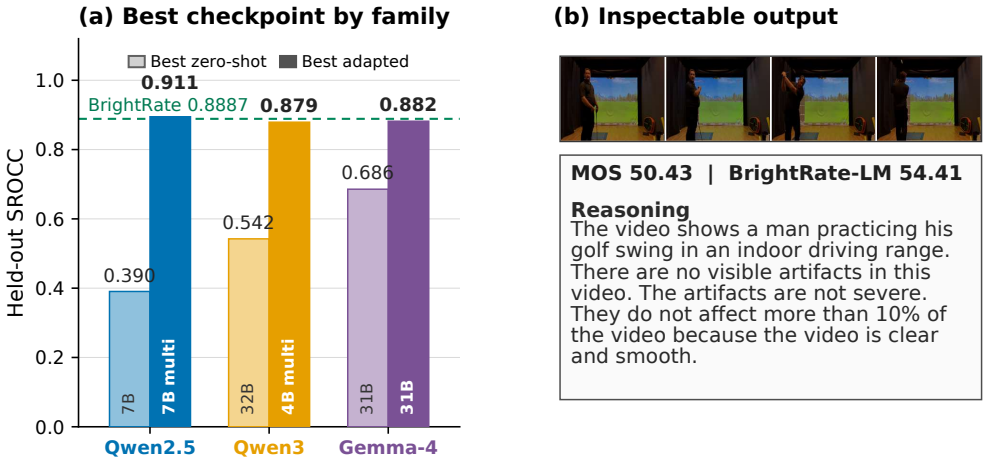


Fig. 1: Adaptation lifts every family, with BrightRate-LM highest. (a) Grouped bars compare the best zero-shot and adapted checkpoints per family; (b) frames accompany the score pair and generated reasoning.

1 Introduction

User-generated high dynamic range (HDR) video combines familiar capture and compression defects with failures concentrated in shadows, highlights, and wide-gamut color. Quality predictors therefore need semantic context and signal evidence that ordinary rendering can hide: more parameters cannot recover discarded evidence, while raw HDR codes can lie outside a pretrained visual distribution. Figure 1 summarizes the result. BrightRate-LM, adapted from Qwen2.5-VL-7B [1], reaches 0.9052 Spearman rank-order correlation coefficient (SROCC) (0.0196) with multi-exposure input, compared with 0.8887 for BrightRate [2], and pairs its score with a defect description and reasoning.

Using BrightVQ [2], we compare generic and quality-specialized zero-shot models, matched adaptation across Qwen3-VL [3] and Gemma-4 [4], and tone-mapped SDR, multi-exposure, and native PQ input. Retained optimization, prediction, transfer, and resource records separate architecture, representation, calibration, and capacity effects. The study makes five contributions:

- Adaptation improves every completed checkpoint but does not make performance monotonic in scale.
- Multi-exposure input lets Qwen3-VL-2B [3] exceed the tone-mapped 8B checkpoint.
- Native-PQ forensics expose two interface bugs; statistics matching recovers 0.8383 SROCC on Gemma-4-12B [4].
- BrightRate-LM jointly provides a score, defect description, reasoning, accuracy, and measured resource use.

- Frozen adapters transfer to Beyond8Bits [5] and CHUG [6], where multi-exposure input is consistently stronger.

Classical no-reference methods model natural image or video statistics [7–12]; learned systems use spatial, semantic, and temporal representations [13–18]. Most target standard dynamic range (SDR) signals despite authentic user-generated mixtures of blur, noise, exposure, motion, editing, and compression [19–25]. HDR10 instead combines PQ, 10-bit precision, and Rec. 2020 primaries [26, 27]. HDR datasets and predictors address this mismatch through perceptual transforms, brightness-aware features, or HDR adaptation [28–38], motivating representation as an experimental variable.

Recent HDR corpora broaden the data regime: CHUG [6] contributes crowdsourced sources and transcodes, while Beyond8Bits [5] provides 28,987 labeled training clips and introduces HDR-Q. Our controlled 2,100-clip study instead holds the benchmark and adaptation recipe fixed while changing the visual interface. Quality-aware multimodal work connects opinion scores to language [39–44]; One-Align [41] supplies the quality-specialized reference. Qwen2.5-VL [1] and Qwen3-VL [3] use dedicated visual encoders, whereas Gemma-4 [4] places an encoder-free 12B pathway among encoder-based family members. These contrasts motivate matched family and interface comparisons instead of a global scaling law.

2 Dataset

We build on the BrightVQ dataset [2], which contains 2,100 HDR10/PQ, Rec. 2020 videos derived from 300 authentic UGC sources. Its 150 landscape and 150 portrait contents each appear at six distorted rate-resolution operating points plus a reference, and 73,794 human ratings are aggregated to MOS with the SUREAL subjective-score model [45]. Each operating point preserves source identity, allowing training and evaluation content to remain separate. Figure 2 shows the breadth of content, lighting, and compression artifacts. Appendix A records the rate-resolution ladder and secondary statistics.

Following BrightRate [2], the classical reference combines two-scale CONTRIQUE features [16], Contrastive Language-Image Pretraining (CLIP) semantics [46], absolute temporal feature differences, and HDR natural-scene statistics after a piecewise expansive luminance transform with $\beta = 4$ and a 31×31 mean-subtracted contrast-normalized (MSCN) support [7, 8, 35]. A radial-basis-function support vector regressor (RBF SVR) selected by five-fold cross-validation maps the normalized clip descriptor to MOS. The recorded BrightVQ result is 0.8887 SROCC and is compared with the language-model results in Section 6. Appendix B points to the full published benchmark, transfer, and ablation evidence.



Fig. 2: BrightVQ spans varied content, lighting, and authentic compression artifacts [2]. The lower examples outline salient artifacts in red.

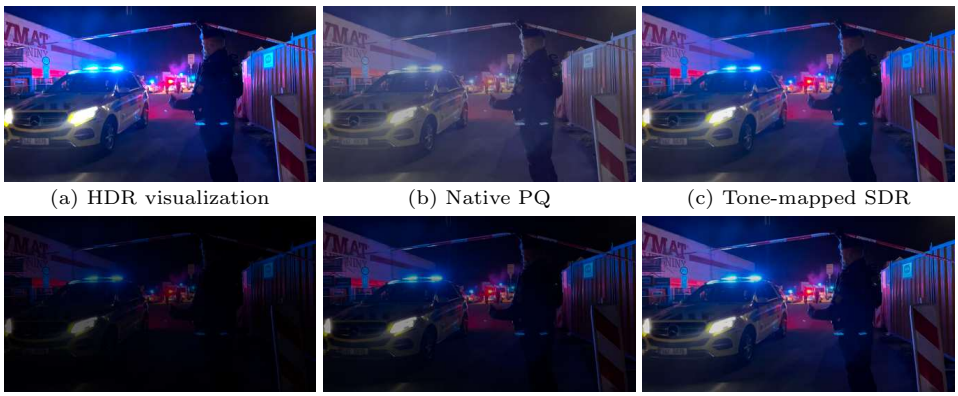
3 Study Setup

The study compares Hable tone-mapped SDR, multi-exposure, and native PQ input; two historical native-PQ variants remain only as defective controls. Tone-mapped input applies the SMPTE ST 2084 EOTF [26], Hable's filmic curve [47], BT.2020-to-BT.709 conversion, and the sRGB transfer, then supplies eight images with maximum edge 448 pixels. The curve uses $(A, B, C, D, E, F) = (0.15, 0.50, 0.10, 0.20, 0.02, 0.30)$, exposure bias 2.0, and white point 11.2. Multi-exposure input applies the same chain at -2, 0, and $+2$ stops, yielding 24 temporal-major images that expose shadows, midtones, and highlights. Native PQ input retains eight truecolor 16-bit BT.2020 frames without tone mapping, gamut conversion, or 8-bit quantization. The corrected Gemma-4-12B interface [4] decodes them to float32 PQ, supplies 255 PQ through the standard processor, and retains its 1/255 rescale, bicubic resize, and learned LayerNorm. It trains `patch_dense` and the vision-to-language projection with the language adapters. Figure 3 compares the interfaces; its HDR and native-PQ tiles are print visualizations, while the model receives the preserved values.[1]

The recovery transforms operate once per eight-frame clip before the corrected processor interface. Let $x_{fchw} \in [0, 1]$ be native PQ and $t_{fchw} \in [0, 1]$ its aligned Hable tone-mapped frames. For channel c, let μ_c and σ_c be population moments over frame and spatial indices. Statistics matching uses

$$x'_{fchw} = \mathrm{clip}_{[0,1]}\left(\frac{x_{fchw} - \mu_c(x)}{\sigma_c(x)}\sigma_c(t) + \mu_c(t)\right). \tag{1}$$

[1] PIL/Pillow misreports these truecolor 16-bit PNGs as 8-bit RGB; the PNG header and a uint16-preserving decoder verify storage precision.



(a) HDR visualization (b) Native PQ (c) Tone-mapped SDR

(d) Exposure −2 stops (e) Exposure 0 stops (f) Exposure +2 stops

Fig. 3: The three model interfaces reveal different evidence from one HDR source. Primary modalities occupy the top row; the −2, 0, and +2 Hable exposures occupy the bottom row.

The reference-free alternative computes scalar quantiles $q_{0.01}(x)$ and $q_{0.99}(x)$ over every frame, channel, and spatial position, then uses

$$x'_{fchw} = \mathrm{clip}_{[0,1]}\left(\frac{x_{fchw} - q_{0.01}(x)}{q_{0.99}(x) - q_{0.01}(x)}\right). \tag{2}$$

Both paths require their denominator to exceed 10^{-6}, then supply $255x'$ to the standard processor. The third completed strategy leaves x unremapped, trains the two visual projections alone for one epoch, and trains them jointly with the language adapters for a second epoch.

Supervised models are trained and evaluated with five 80/20 content-separated repetitions using disjoint source content; their reported values are means, with population standard deviations in parentheses. Zero-shot models and retained reference methods report medians across 100 content-separated partitions, with standard deviations in parentheses. SROCC is the primary rank metric; PLCC, Kendall rank-order correlation coefficient (KRCC), and RMSE are also reported. For zero-shot models only, PLCC and RMSE use an independently fitted four-parameter logistic map within each test set. Every supervised run records step loss, optimizer update, learning rate, peak allocated memory, elapsed wall time, configuration, and checkpoint. Development SROCC is evaluated every 50 updates on a fixed 64-video subset, while held-out predictions retain stable video identifiers and join to MOS and ladder metadata only for analysis. Appendix C gives the completion and aggregation rules.

4 Models

The comparison includes Qwen2.5-VL [1] at 3B and 7B, Qwen3-VL [3] at 2B, 4B, 8B, and 32B, and Gemma-4 [4] at E2B, E4B, 12B, 26B-A4B, and 31B. One-Align [41]

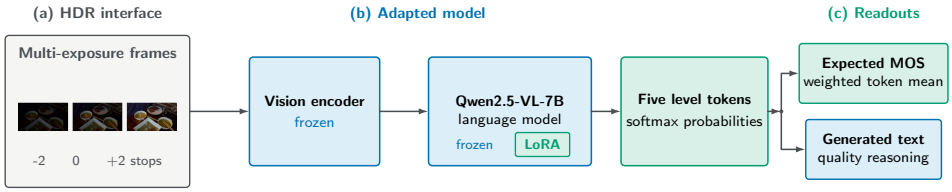


Fig. 4: BrightRate-LM converts HDR evidence into an auditable score and language diagnosis. Direct labels distinguish the input interface, adapted Qwen2.5-VL-7B [1] backbone, and two readout paths.

provides a quality-specialized zero-shot reference. Supervised Hable tone-mapped SDR comparisons cover every listed checkpoint with a completed adapter except Qwen3-VL-32B. Multi-exposure input is compared on paired Qwen2.5-VL-3B and 7B, Qwen3-VL-2B and 4B, and Gemma-4-12B checkpoints. The 72B Qwen3-VL checkpoint was not run because its checkpoint exceeded the available accelerator-memory budget.

Table 1[2] reports every evaluated checkpoint. Qwen2.5-VL uses 32-layer visual transformers; Qwen3-VL uses 24 or 27 layers by scale; and One-Align uses a 24-layer mPLUG-Owl2 visual encoder plus a six-layer abstractor. Gemma-4 alone changes visual architecture within its family: E2B and E4B use 16 layers, 26B-A4B and 31B use 27, and only 12B is encoder-free.

Table 1: Only Gemma-4-12B [4] is encoder-free; every other evaluated checkpoint uses a visual transformer.

Checkpoint	B	Vision path	Pixel path	Adapters (modules)	Evaluation
Qwen2.5-VL-3B [1]	3.755	Q2.5-ViT (32)	3D patch conv.	L+V LoRA (348)	ZS; TM; ME
Qwen2.5-VL-7B [1]	8.292	Q2.5-ViT (32)	3D patch conv.	L+V LoRA (292)	ZS; TM; ME
Qwen3-VL-2B [3]	2.128	Q3-ViT (24)	3D patch conv.	L LoRA (196)	ZS; TM; ME
Qwen3-VL-4B [3]	4.438	Q3-ViT (24)	3D patch conv.	L LoRA (252)	ZS; TM; ME
Qwen3-VL-8B [3]	8.767	Q3-ViT (27)	3D patch conv.	L LoRA (252)	ZS; TM
Qwen3-VL-32B [3]	33.357	Q3-ViT (27)	3D patch conv.	None	ZS
Gemma-4-E2B [4]	5.123	Gemma-ViT (16)	Patch embed.+proj.	L+V+A LoRA (353)	ZS; TM
Gemma-4-E4B [4]	7.996	Gemma-ViT (16)	Patch embed.+proj.	L+V+A LoRA (406)	ZS; TM
Gemma-4-12B [4]	11.960	None	`patch_dense`	L LoRA (328); PQ:+patch+V-to-L	ZS; TM; ME; PQ
Gemma-4-26B-A4B [4]	25.806	Gemma-ViT (27)	Patch embed.+proj.	L+V LoRA (394)	ZS; TM
Gemma-4-31B [4]	31.273	Gemma-ViT (27)	Patch embed.+proj.	L+V LoRA (599)	ZS; TM
One-Align [41]	8.205	mPLUG-ViT (24)+abs. (6)	2D conv.+abstractor	None	ZS

All adapted checkpoints target attention and MLP projections. E2B and E4B require custom dispatch for clippable linear layers. The encoder-free 12B has the fewest Gemma adapters and no pretrained vision transformer; its completed native-PQ interventions additionally train `patch_dense` and the vision-to-language projection. The results separate this architectural outlier from its encoder-equipped siblings.

BrightRate-LM adapts the Qwen2.5-VL-7B-Instruct checkpoint [1] to HDR video quality assessment. Figure 4 separates the signal path, frozen backbone, trainable adapters, and two output routes. The score is a deterministic expectation over five quality-level tokens. Description and reasoning are generated separately by the same adapted model and never alter the numerical readout.

[2]Parameters are totals; 26B-A4B activates 4B per token. L, V, and A denote language, vision, and audio modules; the E2B and E4B visual runs resolve audio projections that their inputs do not activate. ZS, TM, and ME denote zero-shot, tone-mapped SDR, and multi-exposure evaluation.

5 Training

Each clip is sampled at the centers of eight equal temporal intervals. PQ code values are decoded by the SMPTE ST 2084 EOTF [26]; 203 nit reference white maps to 0.18 scene linear at zero stops. Every frame is rendered at -2, 0, and $+2$ stops with the Hable curve, converted from linear BT.2020 to BT.709, clipped to $[0, 1]$, encoded by the sRGB opto-electronic transfer function (OETF), and stored as an 8-bit image with maximum edge 448 pixels. Temporal-major ordering keeps the three exposures adjacent for every time point. For MOS y, the target is a soft distribution over level tokens at knots $m_k \in \{0, 25, 50, 75, 100\}$. Between adjacent knots m_j and m_{j+1}, the two nonzero probabilities are $q_j = (m_{j+1} - y)/25$ and $q_{j+1} = (y - m_j)/25$. Soft cross entropy trains the next-token logits. At evaluation, p_k is the softmax probability of level token k and the model score is $\hat{y} = \sum_k p_k m_k$.

With rank-16 LoRA, BrightRate-LM trains 47,589,376 parameters for 420 optimizer updates. Appendix Table 10 records the complete retained recipe. For the language output, the description prompt asks for visible blockiness, banding, blur, noise, exposure, color, or motion artifacts without mentioning audio or the sampling interface. The reasoning prompt supplies the predicted level word and asks the model to connect visible strengths or defects, their severity, and their spatial or temporal extent to that level. Figure 1 shows an unedited recorded example.

6 Results

6.1 Zero-shot models remain poorly calibrated

Table 2 places BrightRate-LM and the new language-model results beside the retained reference methods. The multi-exposure BrightRate-LM row is best on all four reported metrics. Generic scale helps only gradually: Qwen3-VL [3] rises from 0.4323 at 2B to 0.5424 at 32B in the zero-shot setting, while Gemma-4 [4] reaches 0.6856 at 31B, the strongest generic zero-shot result, but still trails One-Align [41] at 0.7025 and BrightRate [2] at 0.8887. The gap is calibration, not simply checkpoint size.

6.2 Adaptation changes the regime

Figure 5 shows the representative training-loss traces on shared axes. Every trace reduces its soft-label loss, and the synchronized step occurs when epoch 2 begins at update 211. Falling training loss alone does not establish equally successful adaptation: the Gemma-4-12B development trajectory below shows underfitting under the shared recipe. Table 3 gives the corresponding before-and-after scores. Qwen3-VL [3] moves from 0.4323, 0.4940, and 0.5045

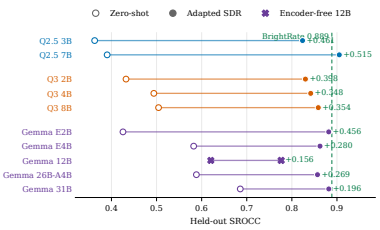


Fig. 6: Zero-shot and adapted SROCC for ten checkpoints; the cross marks the encoder-free 12B.

Table 2: BrightRate-LM with multi-exposure input leads the unified BrightVQ comparison on all four metrics. Dark cells mark column bests.

Class	Model	Visual input	SROCC	PLCC	KRCC	RMSE
IQA	BRISQUE [7]	–	0.3302 (0.0366)	0.3603 (0.0311)	0.2261 (0.0279)	12.58 (0.29)
	HDRMAX [32]	–	0.6276 (0.0321)	0.6318 (0.0356)	0.4409 (0.0288)	10.24 (0.40)
	CONTRIQUE [16]	–	0.7081 (0.0297)	0.7074 (0.0395)	0.5177 (0.0239)	11.46 (1.33)
	REIQA [48]	–	0.7919 (0.0116)	0.8023 (0.0168)	0.6068 (0.0103)	7.94 (0.34)
VQA	V-BLIINDS [9]	–	0.4605 (0.0365)	0.4478 (0.0347)	0.3180 (0.0246)	11.93 (0.42)
	CONVIQT [49]	–	0.7026 (0.0462)	0.7202 (0.0510)	0.5134 (0.0431)	10.58 (1.42)
	VSFA [13]	–	0.7556 (0.0139)	0.7501 (0.0206)	0.5538 (0.0138)	8.83 (0.18)
	COVER [17]	–	0.7609 (0.0201)	0.7917 (0.0252)	0.5597 (0.0181)	7.74 (0.31)
	FasterVQA [15]	–	0.7744 (0.0162)	0.7625 (0.0147)	0.5763 (0.0152)	9.07 (0.25)
	DOVER [18]	–	0.7745 (0.0155)	0.8060 (0.0207)	0.5924 (0.0123)	7.46 (0.28)
	FastVQA [14]	–	0.8094 (0.0121)	0.8530 (0.0156)	0.6445 (0.0106)	7.13 (0.24)
HDR-VQA	HDR-ChipQA [35]	–	0.6781 (0.0220)	0.6855 (0.0179)	0.4889 (0.0160)	9.59 (0.31)
	HIDRO-VQA [36]	–	0.8526 (0.0217)	0.8620 (0.0136)	0.6680 (0.0215)	6.57 (0.34)
	BrightRate [2]	–	0.8887 (0.0197)	0.8970 (0.0171)	0.7059 (0.0227)	5.75 (0.45)
MLLM	Qwen3-VL-32B [3]	Tone-mapped SDR	0.5424 (0.0421)	0.5396 (0.0452)	0.3807 (0.0319)	10.69 (0.51)
	One-Align [41]	Tone-mapped SDR	0.7025 (0.0323)	0.7223 (0.0309)	0.5138 (0.0281)	8.70 (0.41)
	Qwen3-VL-2B [3]	Multi-exposure	0.8704 (0.0221)	0.8803 (0.0137)	0.6845 (0.0290)	6.33 (0.42)
	BrightRate-LM	Tone-mapped SDR	0.8979 (0.0173)	0.9077 (0.0124)	0.7188 (0.0238)	5.65 (0.29)
	BrightRate-LM	Multi-exposure	**0.9052 (0.0196)**	**0.9107 (0.0147)**	**0.7281 (0.0281)**	**5.53 (0.37)**

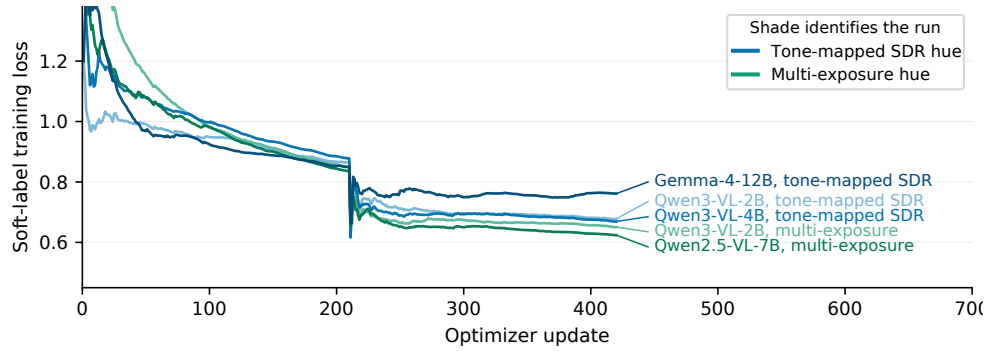


Fig. 5: Training-loss traces are colored by input interface, with tone-mapped SDR and multi-exposure distinguished by color and model variants by shade.

SROCC before adaptation to 0.8301, 0.8418, and 0.8586 after adaptation at 2B, 4B, and 8B. Gemma-4 [4] reverses its before-adaptation ordering and remains non-monotonic across the family, so more parameters do not guarantee a better adapted HDR quality model. The 0.9051 Hable tone-mapped and 0.9110 multi-exposure Qwen2.5-VL-7B entries use the same adapter configurations as the 0.8979 and 0.9052 BrightRate-LM entries in Table 2 under different summaries: Table 3[3] reports one run on a shared held-out content partition, whereas Table 2 reports the mean over five content-separated repetitions.

Table 3 and Fig. 6 show every completed before-and-after pair. The largest gains tend to occur among weaker zero-shot models. Gemma-4-E2B improves by 0.4562 SROCC,

[3]In Table 3, a dash denotes a configuration that was not run, and the Gemma-4-12B adapted value remains the Hable tone-mapped SDR comparison; that checkpoint's best completed result is 0.8383 with native PQ after clip-level statistics matching, a different input interface.

Table 3: Adaptation improves every completed checkpoint without preserving the zero-shot ranking. Values are SROCC on one shared held-out content partition; dark cells mark column bests.

Checkpoint	Params (B)	Zero-shot	Adapted	Multi-exposure
Qwen2.5-VL-3B [1]	3.755	0.3628	$0.8237_{+0.4609}$	0.8875
Qwen2.5-VL-7B [1]	8.292	0.3905	$\mathbf{0.9051}_{+0.5147}$	**0.9110**
Qwen3-VL-2B [3]	2.128	0.4323	$0.8301_{+0.3978}$	0.8606
Qwen3-VL-4B [3]	4.438	0.4940	$0.8418_{+0.3478}$	0.8790
Qwen3-VL-8B [3]	8.767	0.5045	$0.8586_{+0.3541}$	-
Qwen3-VL-32B [3]	33.357	0.5424	-	-
Gemma-4-E2B [4]	5.123	0.4255	$0.8817_{+0.4562}$	-
Gemma-4-E4B [4]	7.996	0.5823	$0.8624_{+0.2801}$	-
Gemma-4-12B [4]	11.960	0.6203	$0.7763_{+0.1560}$	0.6985
Gemma-4-26B-A4B [4]	25.806	0.5883	$0.8568_{+0.2685}$	-
Gemma-4-31B [4]	31.273	**0.6856**	$0.8818_{+0.1962}$	-

while Gemma-4-31B begins with the best generic zero-shot score in its family yet reaches 0.8818 after adaptation, effectively tied with E2B at 0.8817. Thus zero-shot rank does not predict the adapted rank. Qwen3-VL-32B adaptation was not run because of compute cost, and Qwen3-VL-72B was excluded because its checkpoint exceeded the available accelerator-memory budget. Multi-exposure runs were not completed for Qwen3-VL-8B or 32B and Gemma-4 E2B, E4B, 26B-A4B, or 31B because of compute cost. Corrected native-PQ training was attempted only on Gemma-4-12B because the tested injection targets its `patch_dense` projection; it is not directly compatible with the deeper vision paths of the encoder-equipped variants.

Figure 7 compares every predictor on the same 420-video held-out content partition. Per-video classical-baseline predictions are not retained, so the comparison uses the strongest available prediction baseline, One-Align [41], plus adapted Qwen3-VL-2B [3] and Gemma-4-E2B [4] checkpoints.

6.3 Input representation can outweigh scale

Figure 8 shows that Qwen2.5-VL-3B [1] rises from 0.8237 to 0.8875, an absolute gain of 0.0638 SROCC. Qwen3-VL-2B [3] gains 0.0305, reaching 0.8606 and exceeding the 8B checkpoint's 0.8586 with Hable tone-mapped SDR input. The 4B checkpoint gains 0.0372. The intervention is therefore large enough to reverse a fourfold nominal scale comparison.

Gemma-4-12B [4] is the only paired checkpoint for which multi-exposure input reduces performance, from 0.7763 to 0.6985. It is also the only encoder-free checkpoint in the five-model Gemma family. On that same checkpoint, native-PQ statistics matching reaches 0.8383 and percentile stretching reaches 0.8305, gains of 0.0621 and 0.0542 over tone-mapped SDR, while two-stage training reaches 0.7247. The dedicated analysis below separates the observed association from a causal claim. Table 4 gives the complete input-representation comparison.

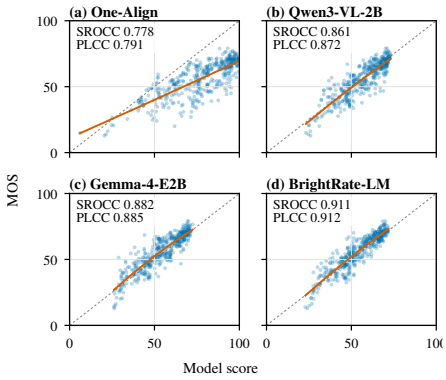


Fig. 7: Correlation plots. BrightRate-LM has the tightest alignment on the shared held-out partition.

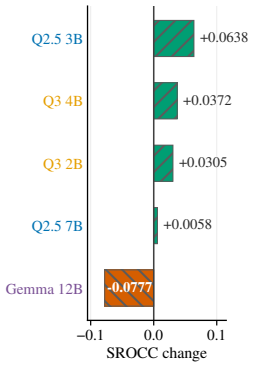


Fig. 8: Multi-exposure helps four of five checkpoints; labels give SROCC changes.

Table 4: Input representation changes ranking more than parameter count in several controlled comparisons. Values are held-out SROCC; the dark cell marks the best result.

Model	Input	SROCC
Qwen2.5-VL-3B [1]	Tone-mapped SDR input	0.8237
Qwen2.5-VL-3B [1]	Multi-exposure input	$0.8875_{+0.0638}$
Qwen2.5-VL-7B [1]	Tone-mapped SDR input	0.9051
Qwen2.5-VL-7B [1]	Multi-exposure input	$\mathbf{0.9110}_{+0.0058}$
Qwen3-VL-2B [3]	Tone-mapped SDR input	0.8301
Qwen3-VL-2B [3]	Multi-exposure input	$0.8606_{+0.0305}$
Qwen3-VL-4B [3]	Tone-mapped SDR input	0.8418
Qwen3-VL-4B [3]	Multi-exposure input	$0.8790_{+0.0372}$
Gemma-4-12B [4]	Tone-mapped SDR input	0.7763
Gemma-4-12B [4]	Multi-exposure input	$0.6985_{-0.0777}$
Gemma-4-12B [4]	Native PQ, clip statistics matched	$0.8383_{+0.0621}$
Gemma-4-12B [4]	Native PQ, percentile stretch	$0.8305_{+0.0542}$
Gemma-4-12B [4]	Native PQ, patch-only then joint	$0.7247_{-0.0515}$

Figure 9 illustrates the quality progression across the ordered encoding settings that underlies the operating-point analysis.

6.4 Native PQ interface recovery

Appendix Figure 17 shows what is established for the corrected native-PQ interface before task adaptation. Supplying 255 PQ through the standard processor retains its rescale, bicubic resize, and learned LayerNorm. At the input to `patch_dense`, native PQ and the Hable tone-mapped reference have means 0.556 and 0.472 and standard deviations 12.27 and 12.23, respectively. Their 1st, 50th, and 99th percentiles are also aligned: $(-31.46, 0.83, 28.22)$ for native PQ and $(-29.53, 0.01, 28.53)$ for the reference. The historical failure therefore cannot be attributed to a marginal scale collapse at the corrected projection input. The recovery experiments target the remaining

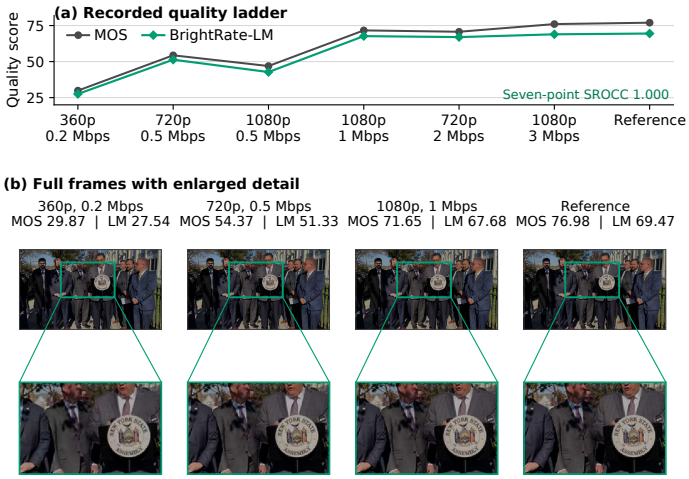


Fig. 9: Compression artifacts recede as bitrate rises, while BrightRate tracks MOS [2]. Labels identify scores, and crops expose local distortion.

Table 5: Corrected native-PQ interfaces recover Gemma-4-12B [4] ranking; historical paths remain as defective controls.

Input and adaptation path	SROCC	PLCC	KRCC	RMSE	Pred. std.
Recovered interfaces					
Tone-mapped, language LoRA only	0.7763	0.7821	0.5833	8.3258	10.1232
Native PQ, clip statistics matched	**0.8383**	0.8505	0.6481	7.1507	12.2037
Native PQ, 1st–99th percentile stretch	0.8305	0.8272	0.6327	7.7162	11.8504
Native PQ, patch-only then joint training	0.7247	0.7442	0.5257	9.3880	12.8506
Defective controls					
Defective native PQ, LayerNorm bypass	0.0100	0.0251	0.0047	13.3218	0.0970
Defective native PQ, second 1/255 rescale	0.0599	−0.0016	0.0456	13.3800	0.1023
Reference dispersion					
Target MOS	–	–	–	–	13.3224

clip-conditioned placement and contrast statistics rather than changing the PQ representation. Table 5 gives the completed recovery results. Per-channel clip statistics matched to the aligned Hable tone-mapped frames reach 0.8383 SROCC with prediction standard deviation 12.2037, exceeding the same checkpoint's 0.7763 tone-mapped result. Mapping each clip's 1st and 99th PQ percentiles to $[0, 1]$ reaches 0.8305 with prediction standard deviation 11.8504 and requires no tone-mapped reference statistics. One patch-only epoch followed by one joint epoch reaches 0.7247 with prediction standard deviation 12.8506. The encoder-free model can therefore learn HDR quality directly from native PQ code values. The decisive requirement is alignment between the statistics reaching the patch projection and those expected by the pretrained projection, not rejection of the PQ representation itself.

The two historical paths remain informative implementation controls. One bypassed the learned LayerNorm. The other retained LayerNorm but applied a second 1/255

rescale, so the model received values 255 times too small. Their prediction standard deviations are 0.0970 and 0.1023, approximately 0.10 in both cases, compared with 10.1 for tone-mapped language-only adaptation and 13.3 for the target MOS. The near-constant predictions establish that these historical failures were implementation artifacts. The corrected plain-rescale run was still in progress at the time of writing. Its tone-mapped patch-training control and the higher-patch-learning-rate probe also remained incomplete, so no final held-out numbers are assigned to them. Native PQ and multi-exposure input were studied only on the anomalous 12B member of Gemma-4. Each native-PQ recovery result is a single completed configuration on this one architecture. The results establish learnability for this encoder-free path but do not estimate variation across repetitions, transfer to encoder-equipped visual stacks, or a family-wide effect for Gemma-4.

6.5 Gemma-4's encoder-free outlier

Gemma-4 [4] adapts strongly: E2B reaches 0.8817 and effectively ties 31B at 0.8818. Table 6 shows that encoder-free 12B has the lowest final and best development SROCC, 0.8402 and 0.8559, and the lowest adapted held-out score, 0.7763. Its native-PQ recovery shows that the shallow path can learn useful ranking when clip statistics are aligned, but architecture and adapter coverage remain correlated, so this result does not generalize to the encoder-equipped variants.

Table 6: Gemma-4 [4] development and held-out correlations; the shaded row is the encoder-free outlier.

Variant	Final dev.	Best dev.	Zero-shot	Adapted
E2B	0.9342	0.9342	0.4255	0.8817
E4B	0.9690	0.9712	0.5823	0.8624
12B	0.8402	0.8559	0.6203	0.7763
26B-A4B	0.9192	0.9192	0.5883	0.8568
31B	0.9274	0.9337	0.6856	0.8818

6.6 BrightRate-LM accuracy and errors

BrightRate-LM, based on Qwen2.5-VL-7B [1], reaches 0.9052 SROCC (0.0196), compared with 0.8979 (0.0173) for its Hable tone-mapped SDR counterpart and 0.8887 (0.0197) for BrightRate [2]. Its PLCC, KRCC, and RMSE are 0.9107, 0.7281, and 5.53. The recorded SROCC values range from 0.8760 to 0.9311, so the gain is not confined to one evaluation partition.

BrightRate-LM reaches 0.911 SROCC and 0.912 PLCC on the displayed prediction set. Figure 10 plots SROCC by operating point, from 0.800 for references to 0.869 at 360p and 0.2 Mbps. Each operating point contains only 60 videos, so these

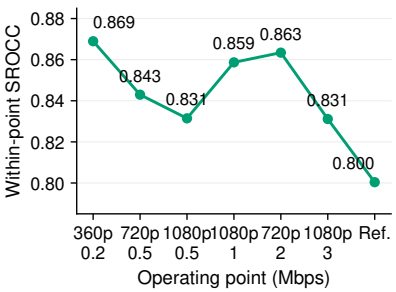


Fig. 10: BrightRate-LM remains robust across operating points, with its lowest SROCC at the reference.

values test robustness rather than establish a precise ordering among settings.

6.7 Adaptation gains by difficulty

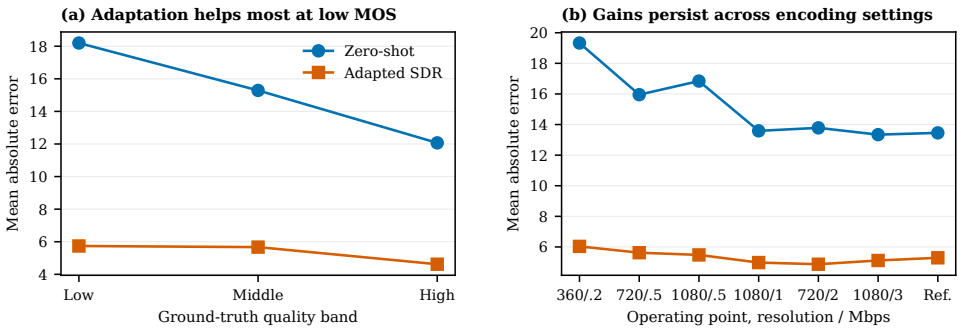


Fig. 11: Adaptation reduces error most for low-MOS videos and the lowest-rate setting. Mean absolute error pools ten paired checkpoints over common held-out videos.

Figure 11 compares zero-shot and adapted predictions for Qwen2.5-VL-3B and 7B [1], Qwen3-VL-2B, 4B, and 8B [3], and all five Gemma-4 checkpoints [4]. In the low-MOS band, adaptation reduces mean absolute error from 18.20 to 5.74, a reduction of 12.46 points. The corresponding reduction is 7.45 points in the high-MOS band. At 360p and 0.2 Mbps, error falls from 19.33 to 6.04, a 13.29-point reduction; at the reference condition it falls by 8.17 points. Resolution and orientation audits agree with this pattern: the largest reduction occurs at 360p, while landscape and portrait videos improve by 10.62 and 8.82 points. Adaptation therefore helps most where generic calibration is weakest, especially at low quality and low bitrate.

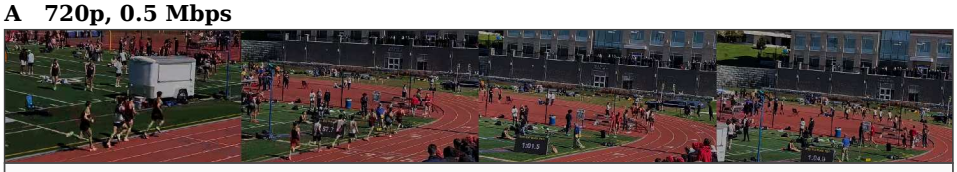
Table 7: Adaptation brings Qwen2.5-VL-7B [1] closer to MOS on difficult examples. Scores use the shared 0-to-100 scale.

`Video ID prefix`	Operating point	MOS	Qwen2.5-VL-7B zero-shot	Adapted Qwen2.5-VL-7B
`9716d357`	360p, 0.2 Mbps	44.82	93.01	46.93
`e2525e24`	1080p, 0.5 Mbps	60.35	80.82	57.89
`9fddfb68`	720p, 2 Mbps	70.97	86.02	69.63

Per-video prediction files for the classical and earlier neural baselines are not present in the retained analysis tree. Consequently, Fig. 11 and Table 7 support only the measurable zero-shot versus adapted language-model comparison. They do not assign per-stratum errors to BRISQUE, CONTRIQUE, DOVER, FastVQA, HIDRO-VQA, BrightRate, or the other benchmark rows.

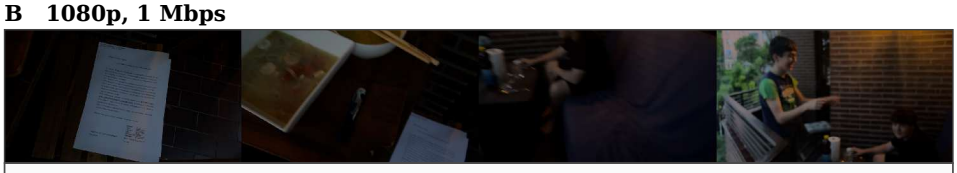
Figure 12 expands the single output card into two held-out examples. In the first, MOS is 31.77 and BrightRate-LM is 31.63; the reasoning connects a daylight track meet with blur, color shift, and motion artifacts. In the second, MOS is 54.84 and BrightRate-LM

A 720p, 0.5 Mbps



Score	MOS 31.77 | BrightRate-LM 31.63
Reasoning	The video captures a track meet with athletes running on a track under bright daylight, but the scene is affected by blurriness and color shifts, with some motion artifacts present as the camera seems to be moving slightly during the recording. Motion artifacts are visible in the video, causing slight blurring and color shifts. These artifacts are moderately severe, affecting the clarity of the image. They impact approximately 50% of the video, which justifies the Poor quality rating due to the significant reduction in visual detail and sharpness.

B 1080p, 1 Mbps



Score	MOS 54.84 | BrightRate-LM 49.24
Reasoning	The video shows two people sitting at a table with bowls of soup and chopsticks. Motion artifacts are visible in the video. They are moderate in severity. They affect most of the video because they are present throughout the entire sequence.

Fig. 12: BrightRate-LM pairs close scores with inspectable language on two held-out videos. Each example shows four frames, the score pair, and generated reasoning.

is 49.24; the reasoning connects a dimly lit table scene with motion artifacts. These examples show that the score and scene-aware reasoning can be inspected together without treating generated language as a guaranteed causal explanation. The qualitative gallery in Appendix C.1 shows every remaining recorded output in MOS order. It includes two material score disagreements: for the nighttime police-checkpoint scene, MOS is 48.79 and BrightRate-LM is 35.16; for the bracelet scene, MOS is 36.41 and BrightRate-LM is 46.26. The high-resolution records also expose the content behind each judgment, including a porch interview, an indoor golf swing, and a podium scene. The gallery therefore shows both close estimates and content-dependent misses without hiding the model's reasoning. Appendix Figure 16 further shows why visible diagnoses are useful when content-dependent score errors remain.

Figure 13 shows that the multi-exposure interface raises peak memory from 4.53 to 5.53 GiB for Qwen3-VL-2B [3] and from 9.13 to 10.29 GiB for the 4B checkpoint. BrightRate-LM uses 19.91 GiB and 6.28 hours in the displayed A100 training record, while its recorded A100 test pass processes 0.40 video/s. The output card in Fig. 1 complements these metrics with the score pair and generated reasoning.

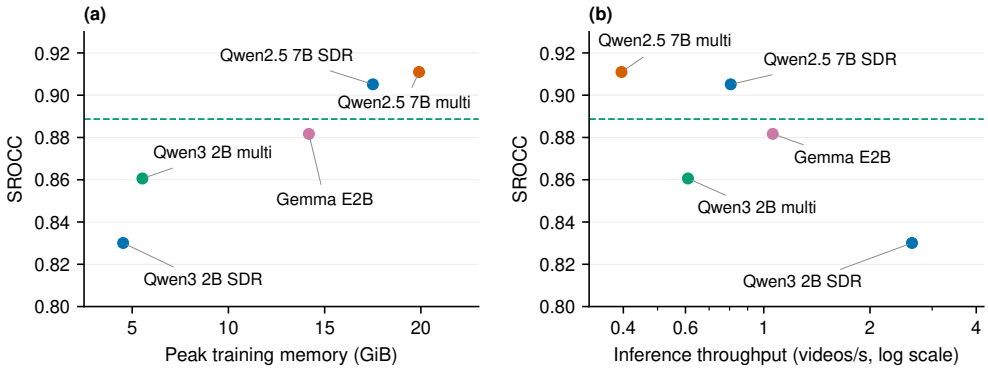


Fig. 13: Resource demand rises with model size and multi-exposure input. Direct labels report peak training memory and inference throughput for five representative models.

6.8 Transfer to larger HDR-UGC corpora

The frozen adapters are evaluated zero-shot after training only on 2,100 BrightVQ clips [2]. On the full 8,281-clip official Beyond8Bits evaluation set [5], BrightRate-LM with multi-exposure input reaches 0.8958 SROCC. It reaches 0.8928 on the 6,706-clip clean subset that excludes clips derived from BrightVQ or CHUG [6]. The full row is therefore not a leakage-free transfer estimate; the clean row measures transfer beyond both predecessor datasets. The tone-mapped interface uses the same Hable rendering as the BrightVQ evaluation. Table 8 reports both input interfaces on the three evaluated sets.

Table 8: Frozen BrightVQ-trained BrightRate-LM transfers to CHUG [6] and Beyond8Bits [5]; dark cells mark block bests.

Dataset	Model	n	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
Tone-mapped SDR						
CHUG present references	BrightRate-LM	856	0.6718	0.7010	0.4813	**5.2453**
Beyond8Bits official test	BrightRate-LM	8,281	**0.8884**	**0.8964**	**0.7056**	5.7057
Beyond8Bits clean test	BrightRate-LM	6,706	0.8839	0.8928	0.6988	5.8222
Multi-exposure						
CHUG present references	BrightRate-LM	856	0.6775	0.7113	0.4874	**5.1698**
Beyond8Bits official test	BrightRate-LM	8,281	**0.8958**	**0.9026**	**0.7160**	5.5417
Beyond8Bits clean test	BrightRate-LM	6,706	0.8928	0.9004	0.7112	5.6250

HDR-Q [5], trained on the 28,987-clip Beyond8Bits training set, reports 0.9206 SROCC in its paper. Its 0.0248 advantage over the transferred multi-exposure adapter on the full official release is useful context, not a head-to-head comparison, because the training data and evaluation procedures differ. The CHUG result has a separate limitation: only 856 source references are present locally, the local store lacks the compressed transcodes, and the available MOS range is compressed. The 0.6775 SROCC therefore describes

reference-only transfer, not performance on the full CHUG bitrate ladder. Across all three evaluated sets, multi-exposure input is stronger than tone-mapped SDR input.

6.9 Classical efficiency baseline

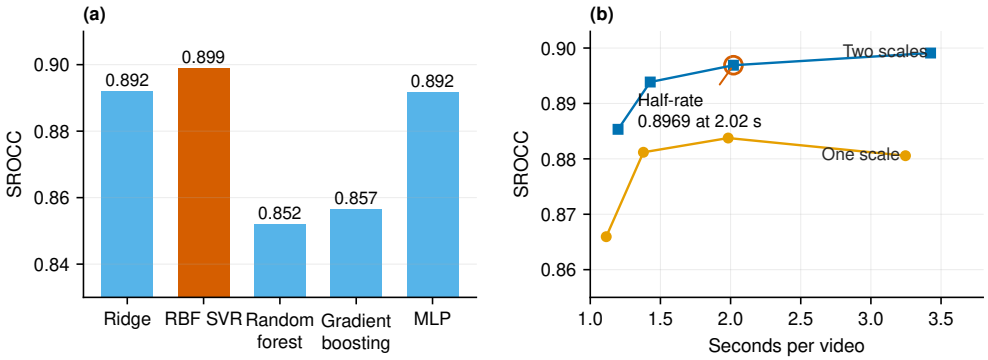


Fig. 14: Controlled extensions of BrightRate [2]. (a) RBF SVR gives the highest median SROCC among five regressors. (b) Half-rate frame sampling retains 0.8969 SROCC at 2.02 seconds per video.

Figure 14 shows the reimplemented 8,740-dimensional BrightRate [2] descriptor evaluated with Ridge, RBF SVR, random forest, gradient boosting, and a two-hidden-layer MLP. RBF SVR is best on every reported metric, reaching 0.8991 SROCC and 5.22 RMSE. The reimplementation differs slightly from the recorded 0.8887 result and is used only for the controlled regressor and sampling analyses. The full two-scale extraction path reaches 0.8991 SROCC at 3.43 seconds per video. Sampling every other frame retains 0.8969 at 2.02 seconds, reducing median wall time by 41 percent for a 0.0022 SROCC difference. Appendix A reports the secondary dataset statistics: portrait mean MOS is 3.010 to 4.365 points higher at every operating point, the constrained SOS relation has negative fit quality, and no frame-rate association survives Holm correction.

7 Discussion

Performance depends on how HDR evidence reaches the pretrained visual interface, not on one family winning everywhere. Multi-exposure input improves every paired Qwen2.5-VL [1] and Qwen3-VL [3] checkpoint, with its largest effect at 3B, so a smaller signal-aware model can outperform a larger tone-mapped model. Scale still matters for the BrightRate-LM backbone, but only conditionally on representation. Gemma-4 [4] also adapts strongly: E2B reaches 0.8817 after a 0.4562 gain and effectively ties 31B at 0.8818. The encoder-free 12B outlier is the only Gemma variant tested with multi-exposure input, so its regression is not a family-wide result.

Native PQ was studied only on Gemma-4-12B [4], whose encoder-free path offers a trainable `patch_dense` injection point. The historical interfaces either bypassed learned LayerNorm or made values 255 times too small through a second rescale; their collapse is an implementation artifact. Correct wiring restores the projection-input distribution, and completed remappings show that clip-conditioned statistics must match the pretrained projection. Statistics matching reaches 0.8383 SROCC, above the 0.7763 Hable tone-mapped result; reference-free percentile stretching reaches 0.8305, while two-stage patch-only then joint training reaches 0.7247. These results establish native-PQ learnability under aligned input statistics, but each recovery is one configuration on one architecture. The tested injection does not directly transfer to the visual transformers in Qwen2.5-VL, Qwen3-VL, or encoder-equipped Gemma-4 variants, which were not run. HDR-Q demonstrates native 10-bit PQ with much larger target-domain training exposure [5]; differing backbones, data, and evaluations preclude a head-to-head inference, but together the results reject an inherent limitation of native PQ.

Adapter coverage also limits cross-family interpretation. The requested suffixes match 292 visual and language modules in Qwen2.5-VL [1] but only 196 language modules in Qwen3-VL [3], whose visual projections use different names. Within-family scale and within-backbone input comparisons remain controlled, but the 7B flagship advantage is not a clean family ranking without a matched visual-module sweep. Repeated supervised evaluations establish repeatability, although a larger adapter campaign would be stronger. Operating-point SROCC uses 60 videos per setting, and the deterministic language example remains qualitative; generated reasoning is diagnostic, not guaranteed causal explanation. Training times span different Lonestar6 GPU paths and are not normalized. The BrightRate [2] reimplementation differs slightly from the original, so its 0.8991 result is confined to the regressor and efficiency controls.

8 Conclusion

This study establishes that generic zero-shot scale is insufficient for HDR video quality assessment. Supervised adaptation changes the performance ordering without producing monotonic scaling, and input representation can outweigh billions of parameters. BrightRate-LM reaches 0.9052 SROCC (0.0196), compared with 0.8887 for BrightRate [2], transfers without retraining at 0.8958 SROCC on the official Beyond8Bits evaluation set [5], and returns an expected MOS with language diagnostics. On Gemma-4-12B [4], two defective historical native-PQ paths are implementation artifacts. Clip-level statistics matching reaches 0.8383 SROCC, exceeding the same checkpoint's 0.7763 Hable tone-mapped result; reference-free percentile stretching reaches 0.8305, and two-stage training reaches 0.7247. The encoder-free model can learn HDR quality from native PQ code values when the statistics entering its pretrained patch projection are aligned. The practical lesson is direct: preserve or statistically align the pretrained visual interface, expose HDR evidence explicitly, and treat scale as one controlled system variable rather than the default route to better prediction. The native-PQ recovery remains limited to single configurations on one architecture.

Acknowledgements. This research was supported by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML); the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for high-performance computing resources; and Google, Inc. for the research collaboration.

Declarations

Funding. Funding was provided by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML).
Competing interests. Shreshth Saini is affiliated with both The University of Texas at Austin and Google, Inc. Yilin Wang, Neil Birkbeck, and Balu Adsumilli are employed by Google, Inc. The research was supported by a collaboration with Google, Inc. The authors have no other competing interests to declare.
Ethics approval and consent to participate. The study analyzes an existing, previously published subjective-quality dataset and collected no new human-subject data for this article. The original crowdsourced study is described in the cited BrightRate article [2].
Data availability. BrightVQ [2] is publicly available at https://shreshthsaini.github.io/BrightVQ/. CHUG [6] and Beyond8Bits [5] are publicly available from their published sources.
Code availability. Code is available at https://github.com/shreshthsaini/BrightRate-LM and the project page https://shreshthsaini.github.io/BrightRate-LM/. Trained adapters are released via the project page.
Author contributions. Shreshth Saini conducted the study, constructed the dataset, designed the models, ran all experiments, and wrote the manuscript. Yilin Wang, Neil Birkbeck, and Balu Adsumilli provided guidance, ideas, and reviews, and Balu Adsumilli additionally reviewed the manuscript in detail. Alan C. Bovik supervised the research.

References

[1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) arXiv:2502.13923 [cs.CV]

[2] Saini, S., Chen, B., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Brightrate: Quality assessment for user-generated hdr videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1522–1532 (2026). https://openaccess.thecvf.com/content/WACV2026/html/Saini_BrightRate_Quality_Assessment_for_User-Generated_HDR_Videos_WACV_2026_paper.html

[3] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H.,

Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) arXiv:2511.21631 [cs.CV]

[4] Gemma Team: Gemma 4 technical report. arXiv preprint arXiv:2607.02770 (2026) arXiv:2607.02770 [cs.CL]

[5] Saini, S., Chen, B., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Seeing beyond 8bits: Subjective and objective quality assessment of hdr-ugc videos. arXiv preprint arXiv:2603.00938 (2026) https://doi.org/10.48550/arXiv.2603.00938 arXiv:2603.00938 [cs.CV]

[6] Saini, S., Bovik, A.C., Birkbeck, N., Wang, Y., Adsumilli, B.: CHUG: Crowdsourced user-generated hdr video quality dataset. In: 2025 IEEE International Conference on Image Processing (ICIP), pp. 2504–2509. IEEE, Anchorage, Alaska, USA (2025). https://doi.org/10.1109/ICIP55913.2025.11084488 . https://doi.org/10.1109/ICIP55913.2025.11084488

[7] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012) https://doi.org/10.1109/TIP.2012.2214050

[8] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013) https://doi.org/10.1109/LSP.2012.2227726

[9] Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Trans. Image Process. **23**(3), 1352–1365 (2014) https://doi.org/10.1109/TIP.2014.2299154

[10] Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Trans. Image Process. **28**(12), 5923–5938 (2019) https://doi.org/10.1109/TIP.2019.2923051

[11] Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Trans. Image Process. **25**(1), 289–300 (2016) https://doi.org/10.1109/TIP.2015.2502725

[12] Ebenezer, J.P., Shang, Z., Wu, Y., Wei, H., Sethuraman, S., Bovik, A.C.: Chipqa: No-reference video quality prediction via space-time chips. IEEE Transactions on Image Processing **30**, 8059–8074 (2021) https://doi.org/10.1109/TIP.2021.3112055

[13] Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: ACM Multimedia, pp. 2351–2359. ACM, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3351028 . https://doi.org/10.1145/3343031.3351028

[14] Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: 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, pp. 538–554. Springer, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20068-7_31 . https://doi.org/10.1007/978-3-031-20068-7_31

[15] Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., Gu, J., Lin, W.: Neighbourhood representative sampling for efficient end-to-end video quality assessment. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 15185–15202 (2023) https://doi.org/10.1109/TPAMI.2023.3319332

[16] Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Image quality assessment using contrastive learning. IEEE Trans. Image Process. **31**, 4149–4161 (2022) https://doi.org/10.1109/TIP.2022.3181496

[17] He, C., Zheng, Q., Zhu, R., Zeng, X., Fan, Y., Tu, Z.: Cover: A comprehensive video quality evaluator. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 5799–5809 (2024). https://doi.org/10.1109/CVPRW63382.2024.

00589 . https://doi.org/10.1109/CVPRW63382.2024.00589

[18] Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20087–20097 (2023). https://doi.org/10.1109/ICCV51070.2023.01843 . https://doi.org/10.1109/ICCV51070.2023.01843

[19] Sinno, Z., Bovik, A.C.: Large-scale study of perceptual video quality. IEEE Trans. Image Process. **28**(2), 612–627 (2019) https://doi.org/10.1109/TIP.2018.2869673

[20] Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The konstanz natural video database (konvid-1k). In: QoMEX, pp. 1–6. IEEE, Piscataway, NJ, USA (2017). https://doi.org/10.1109/QoMEX.2017.7965673 . https://doi.org/10.1109/QoMEX.2017.7965673

[21] Wang, Y., Inguva, S., Adsumilli, B.: Youtube UGC dataset for video compression research. In: MMSP, pp. 1–5. IEEE, Piscataway, NJ, USA (2019). https://doi.org/10.1109/MMSP.2019.8901772 . https://doi.org/10.1109/MMSP.2019.8901772

[22] Ying, Z., Mandal, M., Ghadiyaram, D., Bovik, A.: Patch-vq:'patching up'the video quality problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14014–14024 (2021). https://doi.org/10.1109/CVPR46437.2021.01380 . https://doi.org/10.1109/CVPR46437.2021.01380

[23] Xu, J., Li, J., Zhou, X., Zhou, W., Wang, B., Chen, Z.: Perceptual quality assessment of internet videos. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, pp. 1248–1257. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475486 . https://doi.org/10.1145/3474085.3475486

[24] Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In: El-Saddik, A., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D.P.T., Atrey, P.K., Hossain, M.S. (eds.) Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, pp. 1045–1054. ACM, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3611737 . https://doi.org/10.1145/3581783.3611737

[25] Zhang, Z., Wu, W., Sun, W., Tu, D., Lu, W., Min, X., Chen, Y., Zhai, G.: MD-VQA: Multi-Dimensional Quality Assessment for UGC Live Videos (2023). https://arxiv.org/abs/2303.14933

[26] Society of Motion Picture and Television Engineers: SMPTE ST 2084:2014: High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays (2014). https://pub.smpte.org/latest/st2084/st2084-2014.pdf

[27] International Telecommunication Union: Recommendation ITU-R BT.2020-2: Parameter values for ultra-high definition television systems for production and international programme exchange (2015). https://www.itu.int/rec/R-REC-BT.2020-2-201510-I/en

[28] Azimi, M., Banitalebi-Dehkordi, A., Dong, Y., Pourazad, M.T., Nasiopoulos, P.: Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content (2018). https://arxiv.org/abs/1803.04815

[29] Pan, X., Zhang, J., Wang, S., Wang, S., Zhou, Y., Ding, W., Yang, Y.: Hdr video quality assessment: Perceptual evaluation of compressed hdr video. Journal of Visual Communication and Image Representation **57**, 76–83 (2018) https://doi.org/10.1016/j.jvcir.2018.10.016

[30] Baroncini, V., Mangiatordi, F., Pallotti, E., Agostinelli, M.: Visual assessment of hdr video. In: IS&T International Symposium on Electronic Imaging: Image Quality and System Performance XIII (2016). https://doi.org/10.2352/ISSN.2470-1173.2016.13.IQSP-224 . https://doi.org/10.2352/ISSN.2470-1173.2016.13.IQSP-224

[31] Rerabek, M., Hanhart, P., Korshunov, P., Ebrahimi, T.: Subjective and objective evaluation of hdr video compression. In: 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) (2015)

[32] Shang, Z., Ebenezer, J.P., Venkataramanan, A.K., Wu, Y., Wei, H., Sethuraman, S., Bovik, A.C.: A study of subjective and objective quality assessment of hdr videos. IEEE Transactions on Image Processing **33**, 42–57 (2023) https://doi.org/10.1109/TIP.2023.3333217

[33] Wang, Y., Yim, J.G., Birkbeck, N., Adsumilli, B.: Youtube sfv+ hdr quality dataset. In: 2024 IEEE International Conference on Image Processing (ICIP), pp. 96–102 (2024). https://doi.org/10.1109/ICIP51287.2024.10647927 . IEEE. https://doi.org/10.1109/ICIP51287.2024.10647927

[34] Mantiuk, R.K., Azimi, M.: PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In: Picture Coding Symposium (PCS). IEEE, ??? (2021). https://doi.org/10.1109/PCS50896.2021.9477471

[35] Ebenezer, J.P., Shang, Z., Wu, Y., Wei, H., Sethuraman, S., Bovik, A.C.: Hdr-chipqa: No-reference quality assessment on high dynamic range videos. Signal Processing: Image Communication **129**, 117191 (2024) https://doi.org/10.1016/j.image.2024.117191

[36] Saini, S., Saha, A., Bovik, A.C.: HIDRO-VQA: High dynamic range oracle for video quality assessment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp. 469–479 (2024)

[37] Narwaria, M., Perreira Da Silva, M., Le Callet, P.: Hdr-vqm: An objective quality measure for high dynamic range video. Signal Processing: Image Communication **35**, 46–60 (2015) https://doi.org/10.1016/j.image.2015.04.009

[38] Aamir, N., Mir, J., Nizami, I.F., Shaukat, F., Majid, M.: Hdr-bvqm: High dynamic range blind video quality model. Multimedia Tools and Applications **80**, 27701–27715 (2021)

[39] Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., Lin, W.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. In: International Conference on Learning Representations (2024). https://arxiv.org/abs/2309.14181

[40] Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., Xue, G., Sun, W., Yan, Q., Lin, W.: 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, pp. 25490–25500 (2024)

[41] Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 54015–54029. PMLR, Brookline, MA, USA (2024). https://proceedings.mlr.press/v235/wu24ah.html

[42] You, Z., Cai, X., Gu, J., Xue, T., Dong, C.: Teaching large language models to regress accurate image quality scores using score distribution. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14483–14494 (2025)

[43] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 543–553. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-demo.49 . https://aclanthology.org/2023.emnlp-demo.49/

[44] Zhang, Z., Jia, Z., Wu, H., Li, C., Chen, Z., Zhou, Y., Sun, W., Liu, X., Min, X., Lin, W., Zhai, G.: Q-bench-video: Benchmark the video quality understanding of LMMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3229–3239 (2025)

[45] Li, Z., Bampis, C.G., Janowski, L., Katsavounidis, I.: A simple model for subject behavior in subjective experiments. In: Electronic Imaging, vol. 2020, pp. 131–113114 (2020). https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-131 . https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-131

[46] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR, Brookline, MA, USA (2021)

[47] Hable, J.: Uncharted 2: HDR Lighting. Game Developers Conference. GDC Vault (2010). https://www.gdcvault.com/play/1012351/Uncharted-2-HDR

[48] Saha, A., Mishra, S., Bovik, A.C.: 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, pp. 5846–5855. IEEE, Piscataway, NJ, USA (2023). https://doi.org/10.1109/CVPR52729.2023.00566 . https://doi.org/10.1109/CVPR52729.2023.00566

[49] Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Conviqt: Contrastive video quality estimator. IEEE Transactions on Image Processing **32**, 5138–5152 (2023) https://doi.org/10.1109/TIP.2023.3310344

Appendix A Dataset Details

We build on the BrightVQ dataset introduced with BrightRate [2]. It contains 300 authentic source contents, balanced between portrait and landscape orientation. Each source has six distorted encodes and one reference, yielding 2,100 HDR10/PQ videos in Rec. 2020 color. The 73,794 valid ratings are aggregated with SUREAL [45]. Table 9 lists the seven operating points.

Table 9: BrightVQ uses six distorted rate-resolution settings and one reference per source [2].

Operating point	Resolution	Bitrate
1	360p	0.2 Mbps
2	720p	0.5 Mbps
3	1080p	0.5 Mbps
4	1080p	1 Mbps
5	720p	2 Mbps
6	1080p	3 Mbps
Reference	native	reference

For subject i, video j, and presentation r, SUREAL models the observed score as

$$u_{ijr} = \psi_j + \Delta_i + \nu_i X + \sigma_i Z, \tag{A1}$$

where ψ_j is latent video quality, Δ_i is subject bias, ν_i and σ_i model content inconsistency and subject inconsistency, and $X, Z \sim \mathcal{N}(0,1)$. Equation A1 yields latent quality values that are mapped to the reported 0-to-100 MOS scale.

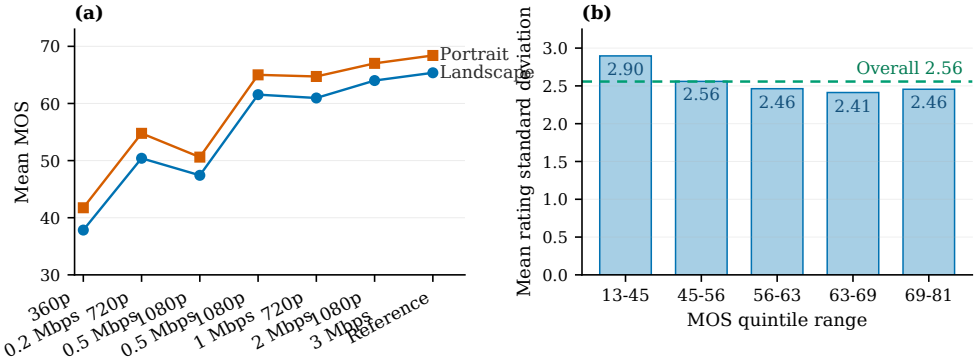


Fig. 15: Portrait MOS is higher at all seven BrightVQ settings, while rating noise is nearly constant [2]. Panels show (a) orientation gaps and (b) binned rating noise.

Figure 15 shows that the orientation shift is significant at every operating point after Holm correction, with adjusted $p \leq 0.0152$, but the landscape and portrait distributions still overlap substantially. The fitted SOS coefficient is $a = 0.0029$, and the fit has

$R^2 = -0.060$. The frame-rate analysis is a useful null result: its minimum adjusted p-value is 0.1711, so no association survives correction.

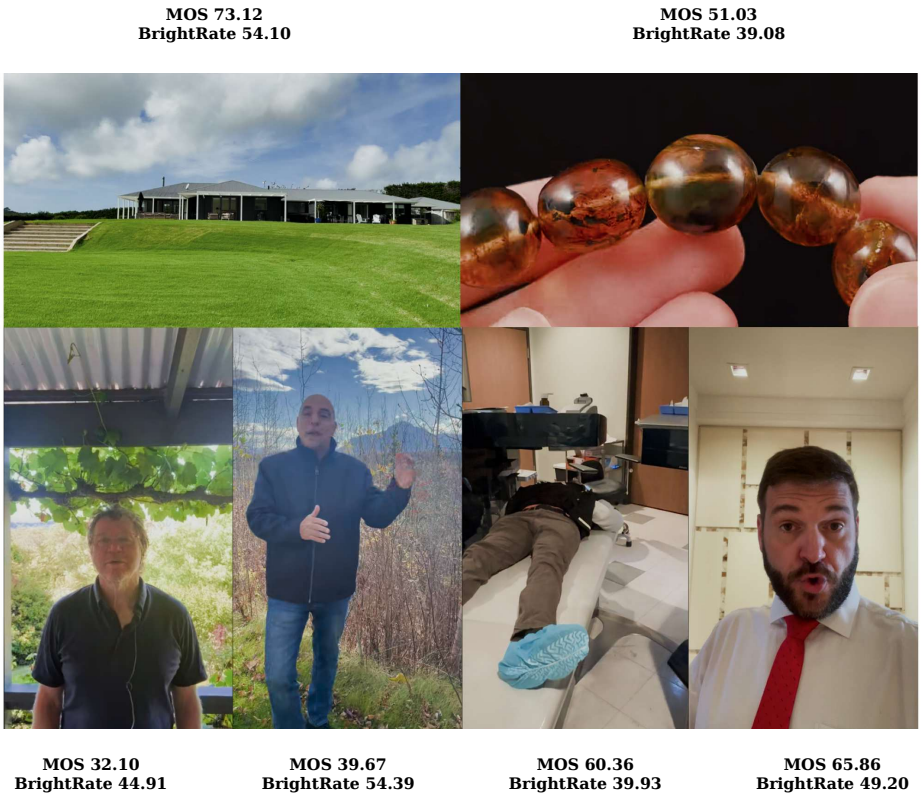


Fig. 16: BrightRate errors remain content dependent across outdoor texture, mixed lighting, and faces [2]. Labels give MOS and model scores on the 0-to-100 scale.

Appendix B Reference Results

The published BrightRate evaluation, including external benchmarks on LIVE-HDR and SFV+HDR, cross-dataset transfer, and branch ablations, is reported in full in the original paper [2]. BrightRate ranks first on both external sets, and its full descriptor outperforms every individual branch and every branch subset.

Appendix C Implementation Details

The input pilot uses the 1,680-video training side of one content partition to select between direct PQ decoding and linear-light Hable rendering followed by BT.709 output.

Tone mapping raises One-Align [41] from 0.5606 to 0.6785 SROCC and Qwen2.5-VL-7B [1] from 0.2950 to 0.3618 on that pilot. The selected path is frozen before the 100 zero-shot test evaluations.

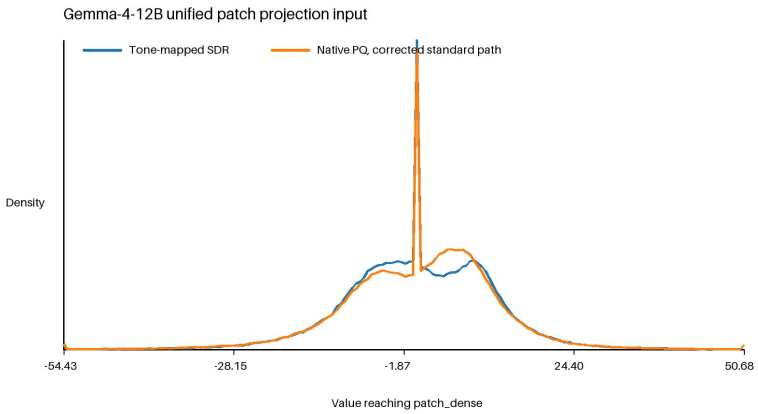


Fig. 17: Correct wiring restores the Gemma-4-12B [4] patch-input distribution of the tone-mapped reference. The diagnostic uses eight test videos before adaptation.

The supervised scripts enforce the same frame identifiers, prompts, five level tokens, soft-label construction, optimizer, and evaluation code across input interfaces. Qwen2.5-VL [1] resolves 292 requested visual and language LoRA modules. Qwen3-VL [3] resolves 196 language modules because its visual projections use different suffixes. The Gemma-4-12B [4] native-PQ interventions additionally train the patch projection and vision-to-language projection. The statistics-matching transform uses per-channel clip means and standard deviations from aligned Hable tone-mapped frames. The percentile transform maps each clip's 1st and 99th PQ percentiles to $[0, 1]$ without reference statistics. The two-stage strategy trains the patch path alone for one epoch and then trains it jointly with the language adapters for one epoch.

A supervised run is complete only when its step log reaches epoch 2 and at least 420 optimizer updates, and a held-out metric file exists for the declared partition. Development curves use a fixed 64-video subset and are never substituted for test metrics. By this rule, statistics matching, percentile stretching, and two-stage training are complete. The corrected plain-rescale interface, tone-mapped patch-training control, and higher-patch-learning-rate probe are incomplete and carry no final metric in this article. Figure generation reads retained JSON, JSONL, and CSV artifacts, validates identifiers and expected row counts, records source-file hashes, and writes vector PDF plus PNG preview pairs. Resource plots use peak recorded allocated memory and completed-run elapsed time. These times are descriptive because the runs used different LS6 GPU paths.

Table 10: The retained BrightRate-LM recipe adapts 47,589,376 parameters with rank-16 LoRA. Settings come directly from completion records and generation scripts.

Component	Recorded setting
Base checkpoint	Qwen2.5-VL-7B-Instruct [1]
LoRA	Rank 16; $\alpha = 32$; dropout 0.05; 47,589,376 trainable parameters (0.5706% of 8,339,756,032 total after adapter insertion)
Target modules	292 resolved `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules
Visual input	8 sampled times; 3 exposures per time; 24 images; maximum edge 448 pixels
HDR rendering	SMPTE ST 2084 PQ to linear light; -2, 0, $+2$ stops; 203 nit $\rightarrow$ 0.18 anchor; Hable filmic operator; BT.2020 to BT.709; sRGB transfer
Labels and loss	MOS interpolation over Bad, Poor, Fair, Good, Excellent at 0, 25, 50, 75, 100; soft cross entropy
Optimization	AdamW; learning rate 10^{-4}; weight decay 0.01; cosine schedule; 5% warmup; gradient-norm cap 1.0; 2 epochs and 420 optimizer updates
Batching	Microbatch 1; gradient accumulation 8; effective batch 8
Hardware and wall time	A100 40 GB: 6.28, 6.00, 6.14, 6.38 h; H100: 3.97 h, one value per training partition
Score readout	Expected MOS under the five level-token softmax, with no fitted test-time calibration
Language readout	Separate visual prompts request a one-to-two-sentence defect description and a two-to-three-sentence justification conditioned on the predicted level; visual evidence only; greedy decoding with one beam; 80 and 160 token caps

C.1 Additional qualitative outputs

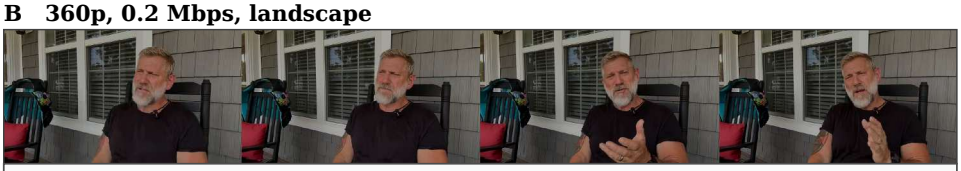
Figures 18–25 contain all ten recorded outputs not used in the main-text qualitative figure. The examples are ordered by MOS and retain every scene-aware reasoning string in full. Each example uses four representative high-resolution frames from the eight-frame sample so the images remain legible at single-column width.

A 1080p, 1 Mbps, landscape



Score	MOS 36.41 | BrightRate-LM 46.26
Reasoning	The video showcases a close-up of a bracelet adorned with a star-shaped pendant and circular elements, set against a plain background. The visible artifacts in the video include color shifts and blurriness around the edges of the bracelet and stars. These artifacts are relatively mild but noticeable enough to detract slightly from the overall clarity. The artifacts affect a significant portion of the video, impacting the perceived quality by making the image less sharp and consistent, thus justifying the Fair level assessment.

B 360p, 0.2 Mbps, landscape



Score	MOS 44.98 | BrightRate-LM 45.94
Reasoning	The man seated on the porch, wearing a dark shirt and a necklace, appears to be engaged in conversation, with his hands gesturing as he speaks. The visible artifacts in the video include blockiness and slight noise, especially noticeable in the darker areas. These issues are relatively mild but still present throughout the video. The artifacts affect a significant portion of the video, impacting the overall clarity and detail, which justifies the Fair quality rating.

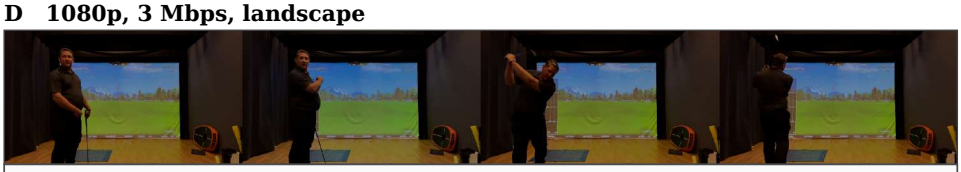
Fig. 18: BrightRate-LM supplies complete, inspectable outputs across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

C 720p, 2 Mbps, landscape



Score	MOS 48.79 | BrightRate-LM 35.16
Reasoning	The video depicts a nighttime scene at a police checkpoint with emergency vehicles and personnel, where the lighting is predominantly blue from the vehicle lights, creating a high-contrast environment with some areas overexposed and others in deep shadow. Motion artifacts are evident throughout the video, causing a slight blur and distortion in the moving elements like the police car and the person standing near it. These artifacts are moderately severe, affecting the clarity and smoothness of the footage. They impact approximately 70% of the video, significantly reducing the overall visual quality and making the scene appear less professional, which justifies the Poor quality assessment.

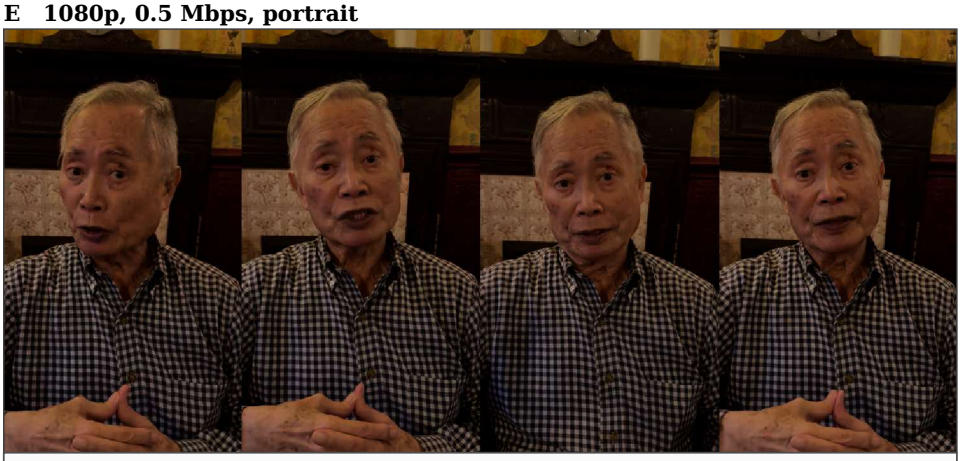
D 1080p, 3 Mbps, landscape



Score	MOS 50.43 | BrightRate-LM 54.41
Reasoning	The video shows a man practicing his golf swing in an indoor driving range. There are no visible artifacts in this video. The artifacts are not severe. They do not affect more than 10% of the video because the video is clear and smooth.

Fig. 19: BrightRate-LM supplies complete, inspectable outputs across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

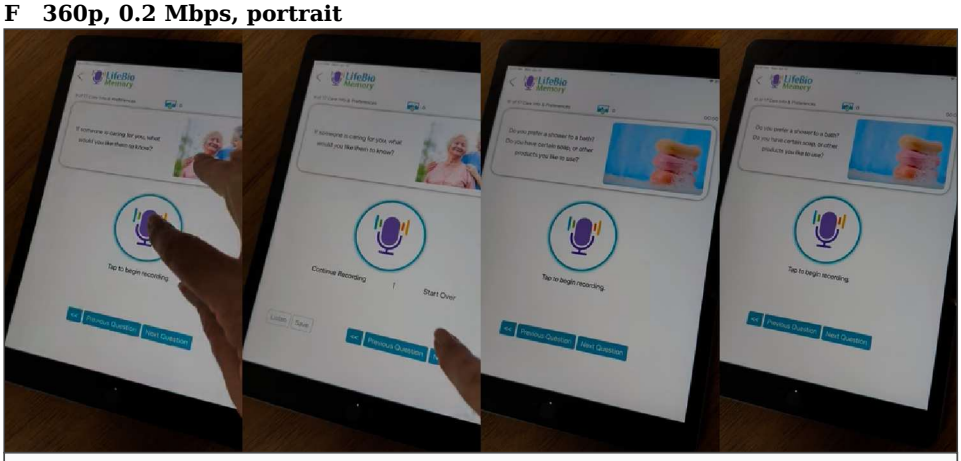
E 1080p, 0.5 Mbps, portrait



Score	MOS 51.34 | BrightRate-LM 55.40
Reasoning	The elderly man in a checkered shirt seated indoors has a relatively dim but sufficient lighting that clearly illuminates his face and upper body. The visible artifacts in the video include noise and minor color shifts. These issues are moderate in severity. They affect approximately half of the video, as the noise is more pronounced in the darker areas, while the color shifts are subtle and scattered.

Fig. 20: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

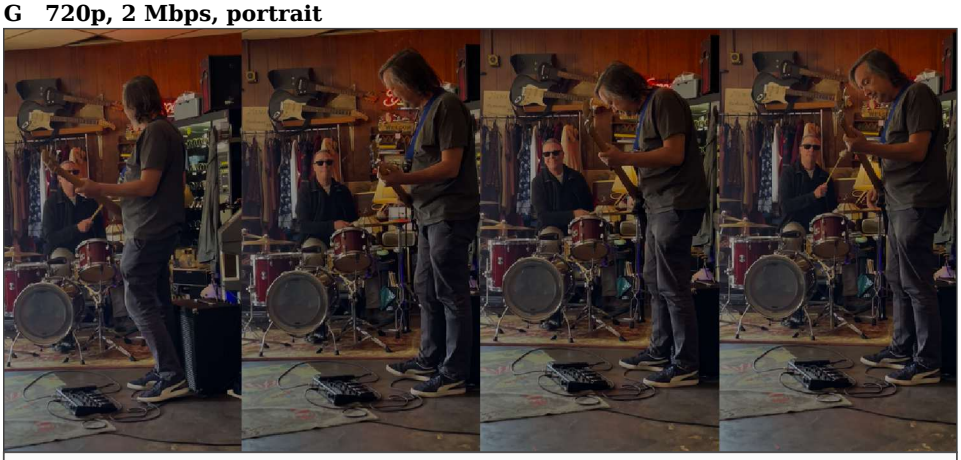
F 360p, 0.2 Mbps, portrait



Score	MOS 55.26 | BrightRate-LM 51.32
Reasoning	The video shows a person interacting with a tablet displaying a LifeBio Memory app on a wooden surface. The visible artifacts in the video are blurriness and slight noise. They are moderate in severity. These artifacts affect a significant portion of the video, as they are present throughout the frames and impact the clarity of the text and images on the tablet screen.

Fig. 21: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

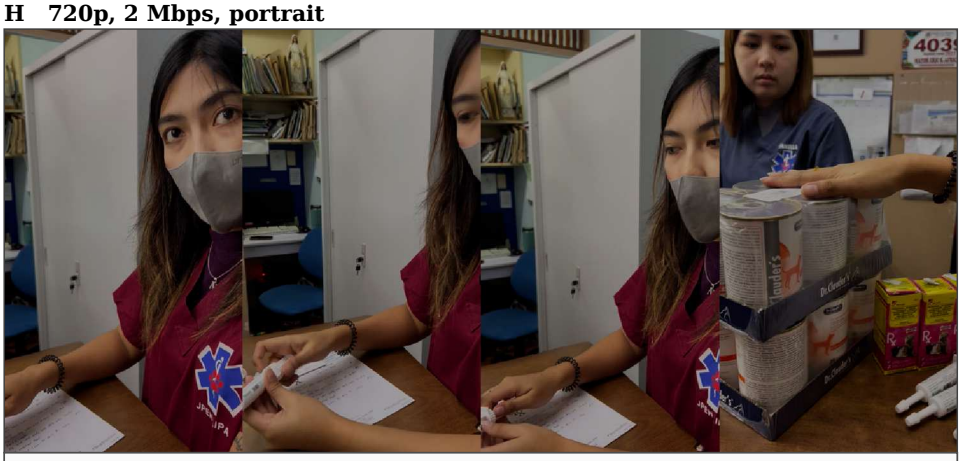
G 720p, 2 Mbps, portrait



Score	MOS 62.53 | BrightRate-LM 57.94
Reasoning	The video captures a man playing guitar in a room filled with musical instruments and equipment. The video has some minor noise present. It is not very severe. It affects most of the video because it slightly detracts from the overall clarity of the image.

Fig. 22: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

H 720p, 2 Mbps, portrait



Score	MOS 69.02 | BrightRate-LM 68.89
Reasoning	The video shows a woman in a medical setting, possibly a clinic or office, as she interacts with various items on a table. There are some color shifts in the video. The color shifts are minor and do not significantly impact the overall quality of the video. This artifact extent justifies the predicted Good level because the video has consistent lighting and no significant changes in brightness or color.

Fig. 23: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

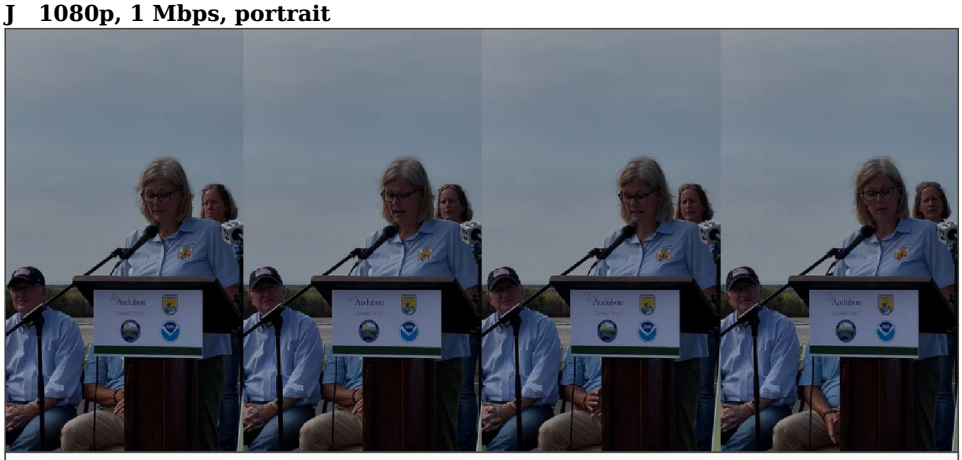
I 720p, 0.5 Mbps, landscape



Score	MOS 70.08 | BrightRate-LM 64.05
Reasoning	The video shows a close-up of a hand with a red heart drawn on it, set against a blurred background of greenery. There are no visible artifacts in the video. None of the mentioned issues are present, as the images appear clear and well-focused. None of the images are affected by any of the mentioned issues, so the video is rated as Good.

Fig. 24: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.

J 1080p, 1 Mbps, portrait



Score	MOS 70.27 | BrightRate-LM 66.29
Reasoning	The video shows a woman speaking at a podium in an outdoor setting. The video has color shifts and some noise. These issues are mild. They affect most of the video but are not severe enough to lower the quality to Poor.

Fig. 25: BrightRate-LM supplies a complete, inspectable output across the held-out MOS range, with frames accompanying the score pair and generated reasoning.