LUMAFLUX: STRUCTURE-AWARE SDR-TO-HDR CONVERSION WITH MONOTONE LUMINANCE FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid adoption of HDR-capable displays has created a growing need to convert legacy and user-generated 8-bit Standard Dynamic Range (SDR) content into perceptually faithful and physically plausible 10-bit High Dynamic Range (HDR). Existing inverse tone-mapping methods often rely on fixed tone mappings or narrowly defined training degradations, which can lead to clipped highlights, chromatic shifts, and temporally unstable tone reproduction. We introduce LUMAFLUX, an Diffusion Transformer (DiT) conditioned directly on physical measurements of the SDR input, while training only 0.57% of total parameters. LUMAFLUX combines Physically-Guided Adaptation (PGA), Perceptual Cross-Modulation (PCM), an HDR Residual Coupler, and a monotone Rational-Quadratic Spline (RQS) tone-field decoder to recover HDR appearance while preserving the luminance ordering of the input. HDR latents are reconstructed from an SDR-conditioned endpoint through direct rectified-flow transport in only 8 steps. The same image model extends to video inverse tone mapping without temporal modules or video-specific training, shared bridge noise and temporally smoothed RQS parameters stabilize deterministic framewise transport. We further curate a 314,396-pair UGC and PGC training corpus in PQ BT.2020 and introduce LumaEval, a unified evaluation protocol spanning eight SDR degradation types. On LumaEval, LUMAFLUX achieves 24.23 dB PSNR and 56.54 ΔE_{ITP}, improving over the strongest baseline by 0.49 dB and 5.06, respectively, while reducing FR-HIDROVQA from 0.703 for HDRTVNet++ to 0.631. For video, temporal stabilization reduces signed excess flicker by 53.3%, while pooled raw flicker remains within 3.3% of the HDR reference. Code will be released.

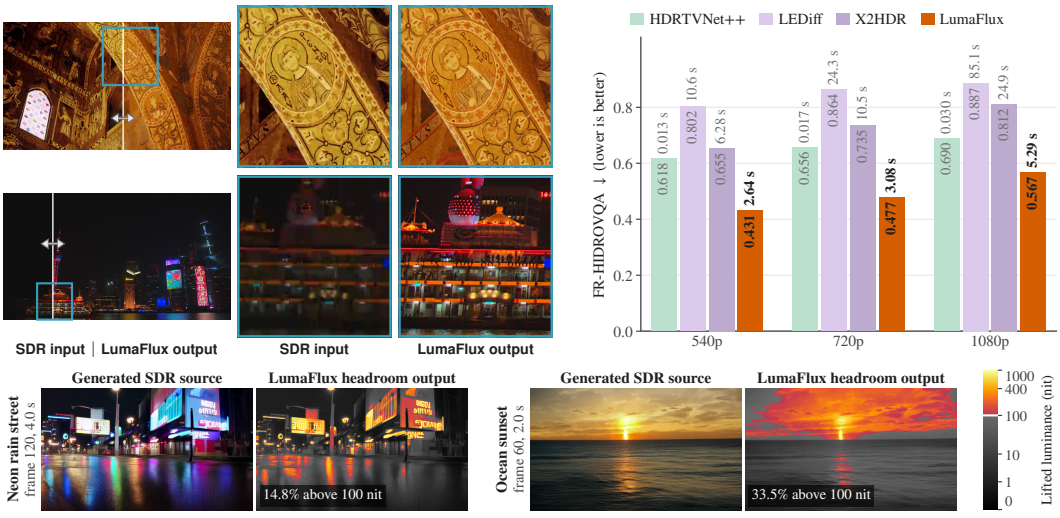

Figure 1: Left: Examples. Right: FR-HIDROVQA bars, seconds per frame above (30 frames). LUMAFLUX scores best at each resolution, 2.4 to 16.1 times faster than X2HDR and LEDiff. Bottom: LUMAFLUX headroom output, two text-to-HDR clips, color above 100 nit (Figure 17).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

High Dynamic Range (HDR) and Wide Color Gamut (WCG) imaging are now standard across modern display and streaming pipelines (ITU-R, 2018; CTA, 2024). ITU-R BT.2020 color primaries (ITU-R, 2015a) and the SMPTE ST 2084 Perceptual Quantizer (PQ) (SMPTE, 2014) define common HDR delivery formats, yet much legacy and user-generated content remains in 8-bit Standard Dynamic Range (SDR) BT.709 (ITU-R, 2015b). Converting such content to PQ BT.2020 enables archival restoration, backward compatibility, and consistent presentation on HDR displays (Figure 1).

Inverse tone mapping (ITM) seeks to recover dynamic and colorimetric range lost during SDR production. An SDR frame $x_{\mathrm{sdr}} \in [0,1]^3$ can be related to an HDR frame x_{hdr} through

$$x_{\mathrm{sdr}} = \Gamma^{709}_{\mathrm{OETF}}\left(\mathcal{M}_{2020\to 709}\left(\frac{\Gamma^{\mathrm{PQ}}_{\mathrm{EOTF}}(x_{\mathrm{hdr}})}{L_{\max}}\right)\right) + \epsilon, \tag{1}$$

where $\Gamma^{709}_{\mathrm{OETF}}$ and $\Gamma^{\mathrm{PQ}}_{\mathrm{EOTF}}$ are the BT.709 opto-electronic and PQ electro-optical transfer functions, $\mathcal{M}_{2020\to 709}$ denotes gamut compression (ITU-R, 2017), $L_{\max}$ is peak display luminance, and ϵ collects quantization and compression error. Inverting this nonlinear, lossy mapping requires priors over both image structure and display-referred luminance. Analytic operators such as Reinhard tone reproduction (Reinhard et al., 2002) and ITU BT.2446 variants (ITU-R, 2021) are efficient, but fixed mappings cannot recover missing highlight structure. Learning-based methods (Eilertsen et al., 2017; Chen et al., 2021b; Kim et al., 2019; Wang et al., 2025b; Li et al., 2024) learn content-dependent mappings, yet often remain tied to a specific synthetic degradation or paired corpus. Generative models provide stronger structural priors, but unconstrained adaptation can introduce semantic, chromatic, or tonal drift.

We introduce LUMAFLUX, a Diffusion Transformer (DiT) (Black Forest Labs, 2024) using measurements of the SDR input (Figure 2). The pretrained prior supplies missing structure, while lightweight trainable paths constrain tone and color. Physically-Guided Adaptation (PGA) gates low-rank attention residuals using luminance, gradient, saturation, and spectral cues, encouraging expansion where supported by the input. Perceptual Cross-Modulation (PCM) injects scene context from frozen SigLIP features (Zhai et al., 2023) through FiLM conditioning (Perez et al., 2018). An HDR Residual Coupler combines both signals across flow time and transformer depth. Because the frozen VAE was trained on SDR imagery, a monotone Rational-Quadratic Spline (RQS) decoder maps its output to a smooth HDR tone field while preserving luminance ordering.

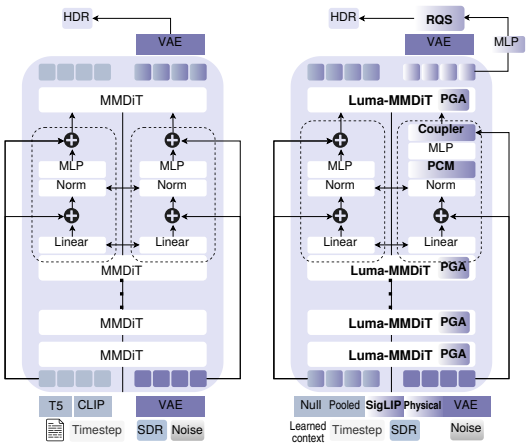

Figure 2: Left: direct MM-DiT adaptation. Right: LUMAFLUX freezes the MM-DiT and VAE, conditions them through PGA, PCM, and the HDR Residual Coupler, then applies the RQS decoder.

To our knowledge, LUMAFLUX is the first display-referred ITM method to steer a frozen pretrained flow transformer using physical measurements of the input in place of text prompts. No text encoder is loaded, and only 71.3M parameters, 0.57% of the loaded weights, are trained. Direct rectified-flow transport reconstructs HDR latents from an SDR-conditioned endpoint in 8 steps. One mixed-corpus checkpoint, without per-track tuning, is best on all seven metrics of the synthetic track (Table 1) and in PSNR and SSIM on expert grades. From QP 27 to QP 42, it loses only 0.55 dB, compared with 2.77 and 2.71 dB for two CNN baselines (Appendix Tables 10 and 13). The same image model transfers to video without temporal layers or video training. Shared bridge noise and smoothed RQS parameters stabilize deterministic framewise transport, reducing signed excess flicker by 53.3%. It also extends to text-to-HDR (Figure 1, bottom), lifting six generated SDR clips to peaks of 739–957 nit (Appendix Tables 8 and 9). Retraining only the adapters on HDRTV1K(Chen et al., 2023) adds 6.65 dB.

Our corpus combines CHUG user-generated content (Saini et al., 2025) with Netflix Open Content (Netflix, 2026), HDRTV1K (Chen et al., 2023), and LIVE-TMHDR (Chen et al., 2021b; Venkataramanan & Bovik, 2024). LumaEval provides a common protocol in which every method uses identical inputs, outputs, and metric implementations across three tracks and matched fixed-QP compression. On its primary track, LUMAFLUX reaches 24.23 dB PSNR and 56.54 ΔE_{ITP}, improving over the strongest baseline by 0.49 dB and 5.06, respectively. Contributions:

- Physical and perceptual adaptation of a frozen diffusion transformer for prompt-free SDR-to-HDR conversion.
- Bottlenecked PGA, PCM, and HDR Residual Coupler paths, a monotone RQS decoder, and 8-step rectified-flow transport.
- A mixed UGC/PGC corpus and deterministic LumaEval protocol with common inputs, metrics, manifests, and content IDs for restricted sources.
- Training-free video ITM using shared bridge noise and smoothed RQS parameters, reducing signed excess flicker by 53.3% without temporal layers, optical flow, or video fine-tuning.

2 Related Work

Analytic ITM expands SDR luminance using prescribed tone curves and saturation rules. Reinhard (Reinhard et al., 2002) and ITU BT.2446 (ITU-R, 2021) are efficient and reproducible, but fixed mappings cannot recover content-dependent structure lost through tone mapping, gamut compression, and quantization. For SDRTV-to-HDRTV conversion, Deep SR-ITM (Kim et al., 2019), HDRTVNet++ (Chen et al., 2023), FMNet (Xu et al., 2022), and KUNet (Wang et al., 2022) learn paired display-referred mappings. Beyond-FMG (He et al., 2025) incorporates real-HDRTV priors, while RealRep (Xu et al., 2026) separates luminance and chrominance across SDR styles. Such models can remain coupled to the degradations represented in their training pairs; LumaEval therefore evaluates synthetic, expert, and native inputs as separate tracks. Video ITM often introduces trained spatiotemporal modules or learned filtering (Xu et al., 2023; Kim et al., 2020), while independently processed frames can propagate input fluctuations into temporally varying HDR mappings.

Among generative approaches, X2HDR (Wu et al., 2026) is most closely related. It freezes the FLUX.1-dev VAE and adapts the denoiser with LoRA in PU21 space for text-to-HDR generation and RAW-to-HDR reconstruction after per-image rescaling to a 4,000-nit peak. LUMAFLUX instead conditions its adapters directly on measurements of the SDR frame and uses a monotone tone decoder for paired SDR restoration in fixed 1,000-nit PQ BT.2020. LEDiff (Wang et al., 2025b) fuses generated exposure latents to recover linear HDR. RB-Modulation (Rout et al., 2025) steers style and content through reference-feature costs and attention aggregation at inference. Our adapters are trained directly for paired HDR fidelity.

3 Preliminaries

A display-referred signal specifies emitted display light in cd/m^2. PQ (SMPTE, 2014) encodes absolute luminance up to 10,000 cd/m^2 into $[0, 1]$, while BT.2020 (ITU-R, 2015a) defines the color primaries. Our 1,000 cd/m^2 mastering peak corresponds to PQ code ≈ 0.752. We distinguish luminance, computed from linear-light RGB and representing physical intensity, from luma, computed from PQ-coded components. The adapters operate on luminance, while the tone decoder acts on luma. Given an HDR frame x_{hdr} in display-referred luminance and wide-gamut chromaticity, the production pipeline maps it to an SDR frame x_{sdr} through Equation 1. ITM estimates $\widehat{x}_{\mathrm{hdr}} = f_{\boldsymbol{\theta}}(x_{\mathrm{sdr}})$ in PQ BT.2020. This inverse is ill-posed because multiple HDR signals can produce the same quantized SDR observation. We therefore combine physical luminance cues with the structural prior of a pretrained generative model.

Flux (Black Forest Labs, 2024) and Stable Diffusion 3 (Esser et al., 2024) use multimodal diffusion transformers (MM-DiT) with flow-based dynamics and attention. A velocity network $\mathcal{F}_{\boldsymbol{\theta}}$ evolves latent state $\mathbf{z}_t$ under conditioning $\mathbf{c}$ as

$$\frac{d\mathbf{z}_t}{dt} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c}), \tag{2}$$

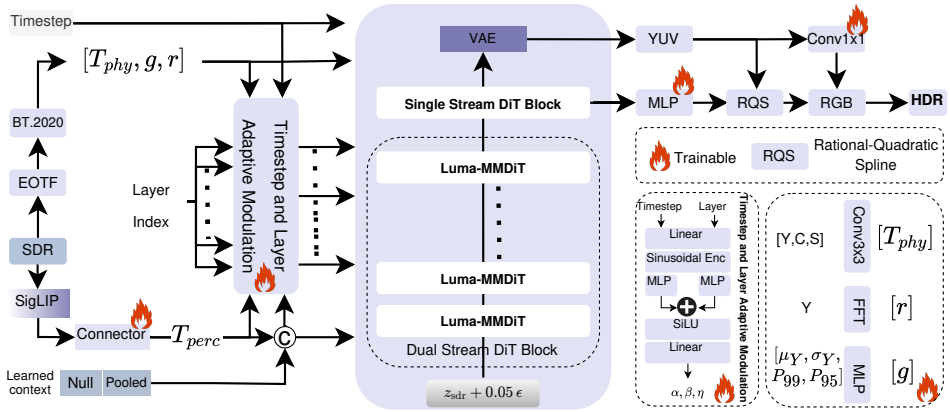

Figure 3: In LUMAFLUX, the SDR input feeds a physical stream and a SigLIP stream. Timestep and layer conditioning controls PGA, PCM, and the HDR Residual Coupler inside each frozen MM-DiT block; a frozen VAE and the monotone RQS tone field produce PQ BT.2020 output. $C = \log(1 + |\nabla Y|)$, S is saturation, and the null and pooled contexts are learned.

using flow matching (Lipman et al., 2023) in the rectified-flow formulation (Liu et al., 2023). We repurpose this transport as a bridge between encoded SDR and HDR images. During training, $\mathbf{z}_t = (1-t)\mathbf{z}_{\text{hdr}} + t\mathbf{z}_1$, $t \in [0,1]$, interpolates between the VAE-encoded HDR target $\mathbf{z}_{\text{hdr}}$ and lightly noised SDR endpoint $\mathbf{z}_1$ from Equation 13; the network predicts the straight-path velocity $\mathbf{z}_1 - \mathbf{z}_{\text{hdr}}$. At inference, integration begins at $\mathbf{z}_1$ and proceeds toward $t = 0$, with each step adding $(t_{k+1} - t_k)\mathcal{F}_{\boldsymbol{\theta}}$ for $t_{k+1} < t_k$, yielding the HDR estimate $\mathbf{z}_0$. Each MM-DiT layer ℓ applies attention and feed-forward residuals to N image tokens of width d, $\mathbf{z}^\ell \in \mathbb{R}^{N \times d}$. LUMAFLUX freezes these weights and modifies only their hidden states through lightweight adapters.

4 METHOD

LUMAFLUX adapts a frozen MM-DiT (Black Forest Labs, 2024) for SDR-to-HDR reconstruction. Figure 3 shows its four trainable components, PGA, PCM, the HDR Residual Coupler, and the RQS tone-field decoder. We wrap all 57 backbone blocks, comprising 19 dual-stream and 38 single-stream blocks ($d = 3072$, 24 heads), PCM precedes each frozen block, PGA modifies its attention, and the coupler follows it. The trainable path contains 71M parameters, only 0.57% of the 12.48B total weights (Appendix Table 17).

4.1 FEATURE EXTRACTION AND ADAPTIVE CONDITIONING

We convert the SDR input to linear-light BT.2020, x_{lin}, and compute luminance $Y = \mathbf{m}_{2020}^\top x_{\text{lin}}$ with $\mathbf{m}_{2020} = [0.2627, 0.6780, 0.0593]^\top$. Luminance, log-gradient, and saturation indicate bright, structured, and colorful regions. A convolution layer converts them to one physical descriptor per image token:

$$T_{\text{phys}} = \text{Conv}_{3\times3}([Y, \log(1 + |\nabla Y|), \text{sat}])\,. \tag{3}$$

Global exposure statistics $s_g = [\mu_Y, \sigma_Y, p_{95}, p_{99}]$ are embedded as $g = \text{MLP}_g(s_g) \in \mathbb{R}^{16}$. A frozen SigLIP-SO400M encoder (Zhai et al., 2023) at 384^2 produces $N_p = 729$ perceptual tokens, $T_{\text{perc}} = \phi_{\text{SigLIP}}(x_{\text{sdr}}) \in \mathbb{R}^{N_p \times 1152}$. A width-64 connector C_{perc} projects them to the MM-DiT width d. No text encoder is used, the text stream instead receives 8 learned null tokens and the connected SigLIP tokens, while the pooled-prompt vector is a learned constant plus a zero-initialized map of g. We also pool a real-valued 2D Fourier transform of Y into $K_r = 8$ radial bands. Their log mean energies form $r \in \mathbb{R}^{K_r}$, summarizing the high-frequency structure that survives tone mapping and compression, a cue also used by FMNet (Xu et al., 2022).

For timestep- and layer-adaptive modulation, a shared conditioner $\Psi(t, \ell)$ produces

$$[\alpha^{t,\ell}_{\text{pga}}, \beta^{t,\ell}_{\text{pga}}, \alpha^{t,\ell}_{\text{pcm}}, \beta^{t,\ell}_{\text{pcm}}, n^{t,\ell}_{\text{spec}}, \lambda^{t,\ell}] = \Psi(t, \ell). \tag{4}$$

The α and β terms control PGA and PCM, n_{spec} weights the spectral gate, and λ gates the coupler. Ψ combines sinusoidal flow-time features with learned block embeddings and SiLU affine heads.

Sigmoid bounds α and λ, while softplus keeps $n_{\text{spec}} \geq 0$. Initialization gives $\alpha = \lambda \approx 0.12$ and $n_{\text{spec}} \approx 0.02$, yielding weak but active adapter paths. Ψ has 41,094 parameters; replacing it with six learned constants costs 0.81 dB and 6.09 ΔE_{ITP} (Table 3).

4.2 Physical and Perceptual Adaptation

Physically-Guided Adaptation (PGA). PGA injects luminance and spectral evidence into attention through gated low-rank residuals related to LoRA (Hu et al., 2022). For frozen value projection $\mathbf{W}_V^{(0)}$, it learns $R_v^{\text{base}} = \mathbf{A}_v \mathbf{B}_v \in \mathbb{R}^{d \times d}$, with $\text{rank}(R_v^{\text{base}}) \leq r_{\text{adapt}} \ll d$. Two per-head gates determine where this residual acts:

$$\mathbf{G}_{\text{phys},i} = \text{Diag}_d(\sigma(P_v[T_{\text{phys},i} \| g])), \qquad g_{\text{FFT}} = \text{softplus}(W_r r). \tag{5}$$

Diag_d repeats each head gate over its channels. The physical gate varies spatially across token i, whereas the spectral gate shares frame-level frequency evidence. Equation 4 combines them as

$$R_{v,i}^{t,\ell} = \left(\alpha_{\text{pga}}^{t,\ell} R_v^{\text{base}} + \beta_{\text{pga}}^{t,\ell} \mathbf{I}\right) \mathbf{G}_{\text{phys},i} \left(\mathbf{I} + n_{\text{spec}}^{t,\ell} \text{Diag}_d(g_{\text{FFT}})\right). \tag{6}$$

Token i therefore uses $\mathbf{W}_V^{(0)} + R_{v,i}^{t,\ell}$ while the frozen projection remains unchanged. We use $r_{\text{adapt}} = 8$ and modify only the image-token value projection. $\mathbf{B}_v$ and W_r start at zero.

Perceptual Cross-Modulation (PCM). PCM conditions normalized transformer states on SigLIP features through FiLM (Perez et al., 2018). Let $T'_{\text{perc}} = C_{\text{perc}}(T_{\text{perc}})$ and $\overline{T}'_{\text{perc}} = N_p^{-1} \sum_i T'_{\text{perc},i}$. For hidden state $\mathbf{h}^\ell$,

$$[\gamma^{t,\ell}, \zeta^{t,\ell}] = \alpha_{\text{pcm}}^{t,\ell} \text{MLP}(\overline{T}'_{\text{perc}}) + \beta_{\text{pcm}}^{t,\ell}, \tag{7}$$

$$\text{PCM}(\mathbf{h}^\ell) = \mathbf{h}^\ell + \gamma^{t,\ell} \odot \text{LN}(\mathbf{h}^\ell) + \zeta^{t,\ell}. \tag{8}$$

The pooled embedding supplies one scale and shift per frame, broadcast across image tokens, while spatial detail is handled by the coupler. Zero initialization of γ and ζ preserves the pretrained representation at the start of training.

HDR Residual Coupler. The coupler injects both conditioning streams at spatial resolution:

$$\mathbf{z}_{\text{out}}^\ell = \mathbf{z}_{\text{res}}^\ell + \lambda^{t,\ell} \left[W_p(T_{\text{phys}}) + W_c(\mathcal{R}(C_{\text{perc}}(T_{\text{perc}})))\right]. \tag{9}$$

Here $\mathbf{z}_{\text{res}}^\ell$ is the frozen MM-DiT output, $\mathcal{R}$ resamples the N_p SigLIP tokens to the latent grid, and W_p, W_c are pointwise projections with width-64 bottlenecks. Their final layers and the PCM output are zero-initialized, so training starts from the frozen network under the new context.

4.3 Rational-Quadratic Spline Tone-Field Decoder

The VAE is trained on SDR images, so its decoded luminance is not itself a calibrated HDR signal. We therefore apply a monotone RQS (Durkan et al., 2019). The frozen decoder produces $x_{\text{out}} = \mathcal{D}_{\text{VAE}}(\mathbf{z}_0)$, interpreted as PQ BT.2020 and converted to Y′CbCr. A head on the pooled final latent predicts spline widths, heights, and slopes $(\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{s})$, while a pointwise projection refines chroma:

$$\widehat{Y} = \text{RQS}(Y_{\text{out}}; \boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{s}), \tag{10}$$

$$\widehat{x}_{\text{hdr}} = \mathcal{M}_{\text{YUV}\to\text{RGB}}([\widehat{Y}, \widehat{U}, \widehat{V}])_{\text{PQ,BT.2020}}. \tag{11}$$

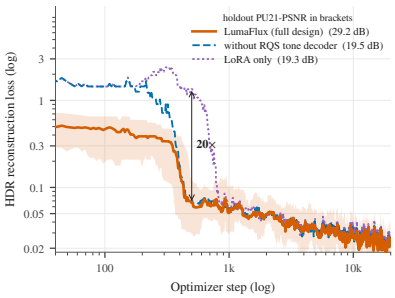

Figure 4: **Optimization dynamics.** Full LumaFlux reaches 29.2 dB; without RQS and with LoRA only, it reaches 19.5 and 19.3 dB.

The spline operates on luma normalized by the 1,000-nit PQ code. Positive widths, heights, and slopes make the mapping differentiable, invertible, and monotone, preserving decoded luma ordering while controlling tone expansion and the highlight knee. We use 8 bins, giving 25 parameters per frame. Removing the RQS head costs 9.69 dB (Table 3; Figure 4).

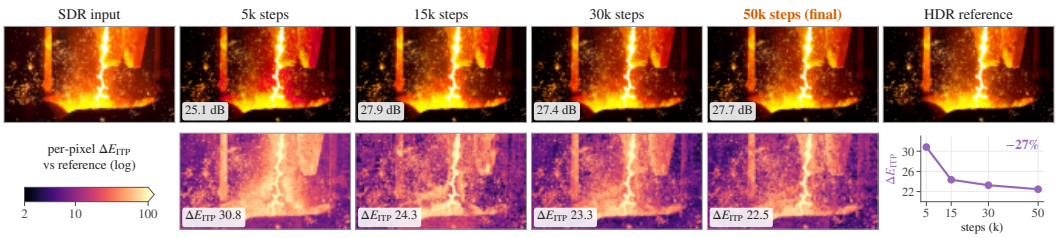

Figure 5: **Training evolution.** HDR reconstruction improves from 5k to 50k steps, with decreasing per-pixel ΔE_{ITP} relative to the HDR reference.

4.4 Training and Video Inference

Given PQ BT.2020 reference $x^\star$ and prediction $\widehat{x}$, we augment rectified-flow velocity matching with decoded reconstruction terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_1 \|Y_{\mathrm{lin}} - Y^\star_{\mathrm{lin}}\|_1 + \lambda_2 \|x_{\mathrm{lin}} - x^\star_{\mathrm{lin}}\|_1 + \lambda_3 \mathcal{L}_{\text{spline-smooth}}. \tag{12}$$

$\mathcal{L}_{\mathrm{FM}}$ regresses the bridge velocity defined in Section 3; linear-light values are obtained through inverse PQ relative to the 1,000-nit peak, and the spline term regularizes adjacent knot slopes. We use $(\lambda_1, \lambda_2, \lambda_3) = (1, 0.5, 0.1)$. Decoded losses are evaluated every fourth optimizer step on the one-step estimate $\widehat{\mathbf{z}}_0 = \mathbf{z}_t - t\,\widehat{\mathbf{v}}_t$, avoiding sampler unrolling. Figure 5 shows the corresponding visual reconstruction trajectory. At inference, if $\mathbf{z}_{\mathrm{sdr}}$ is the frozen-VAE encoding of the SDR frame, we initialize

$$\mathbf{z}_1 = \mathbf{z}_{\mathrm{sdr}} + 0.05\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{13}$$

then integrate from $t = 1$ to $t = 0$ in $K = 8$ uniform Euler steps with fixed guidance. A 1080p frame requires 5.28 s and 27.06 GB, compared with 24.99 s at $K = 40$ with no calibration gain (Appendix J). No partial-noise start or sampling-strength parameter is used. LumaFlux extends the image model directly to HDR video, without temporal layers, optical flow, video fine-tuning, or temporal loss. Fixing bridge noise $\boldsymbol{\epsilon}$ and the solver makes framewise transport deterministic. Reusing the same $\boldsymbol{\epsilon}$ across consecutive frames removes independent sampling variation and encourages nearby inputs to follow corresponding trajectories. The RQS head predicts $\boldsymbol{\theta}_f = (\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{s})$ for frame f, but applies $\overline{\boldsymbol{\theta}}_f = 0.8\overline{\boldsymbol{\theta}}_{f-1} + 0.2\boldsymbol{\theta}_f$. This prior-frame EMA suppresses jitter in global exposure and the highlight knee. Neither intervention alters training, model weights, or architecture; together they reduce signed excess PU21 flicker by 53.3% (Appendix Table 8, Figure 15).

5 Corpus and Evaluation Data

The corpus contains 314,396 SDR-HDR pairs from 2,092 source videos: 856 CHUG references (846 for training and ten reserved), 40 LIVE-TMHDR references with 120 expert SDR clips, 1,235 native HDRTV1K pairs, and Netflix's *Sol Levante* HDR10 master (Saini et al., 2025; Venkataramanan & Bovik, 2024; Chen et al., 2021b; Netflix, 2026). Training draws from 171,384 UGC and 143,012 PGC examples. As shown in Figure 6, the training frames span both benchmark distributions in frame statistics and broadly cover Rec.2020 chromaticities. Eight tone-mapping operators, gamut conversion, and three CRF settings generate missing SDR counterparts, while all HDR references are normalized to PQ BT.2020. Appendix O records source identities, degradation parameters, and checksums for deterministic reconstruction without redistributing restricted media. LumaEval's primary synthetic track fixes 100 aligned frame IDs from 20 held-out contents and macro-averages across 78 video-variant groups. Each method is scored only

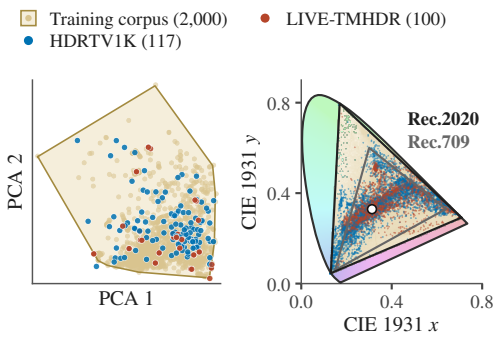

Figure 6: Corpus diversity. (a) PCA of ten frame statistics with the training-corpus hull. (b) CIE xy chromaticity footprint of the corpus (sand) relative to evaluation benchmarks.

Table 1: LumaEval macro-averages over the same 100 IDs and 78 video-variant groups; methods with incomplete coverage are omitted. Bold and underline mark best and second best.

	PU21			Color	Image	HDR	Video
Method	PSNR↑	PSNR-Y↑	SSIM↑	ΔE_{ITP} ↓	HDR-LPIPS↓	HDR-VDP-3↑	FR-HIDRO↓
BT.2446c	<u>23.74</u>	24.22	<u>0.8291</u>	<u>61.60</u>	0.307	5.611	0.753
Reinhard operator	23.62	<u>24.25</u>	0.8286	61.76	<u>0.295</u>	<u>5.726</u>	0.753
HDRTVNet++	23.03	23.23	0.8192	66.67	0.328	5.417	<u>0.703</u>
FMNet	22.86	22.97	0.8176	67.92	0.330	5.412	0.724
ITM-LUT	22.92	23.06	0.8142	67.41	0.315	5.496	0.761
VAE + RQS (control)	18.54	19.41	0.7647	111.67	0.381	3.673	0.924
LUMAFLUX	**24.23**	**24.57**	**0.8294**	**56.54**	**0.269**	**5.812**	**0.631**

on these aligned IDs, which excludes KUNet, PromptIR, LEDiff, and X2HDR from Table 1 (Appendix L). The expert and native tracks contain 100 and 117 pairs, respectively. All outputs are evaluated as PNG16 PQ BT.2020 at 1,000 nit.

6 EXPERIMENTS

Our scripts produce every LumaEval row under a common scoring stack. The analytic rows invert BT.2446 Method C and Reinhard under the fixed mastering convention. LEDiff and X2HDR predict relative radiance, so a common wrapper anchors their 99.5th luminance percentile at 1,000 nit, converts BT.709 to BT.2020, clips at the peak, and applies ST 2084. PromptIR (Potlapalli et al., 2023) serves as a restoration control followed by a fixed Reinhard inverse; learned baselines are HDRTVNet++, FMNet, and ITM-LUT (Guo et al., 2023b).

The MMDiT backbone remains frozen. PGA, PCM, the coupler, conditioner Ψ, and RQS head train on 512^2 crops ($N = 1{,}024$ tokens) for 100,000 steps with global batch size 16 and learning rate 2×10^{-4}, requiring about 36 hours (144 GPU-hours) on four GH200 GPUs (one per node). Prompt-free inference integrates Equation 13 in 8 steps (runtime in Appendix J). We report RGB and luminance PU21-PSNR, luminance PU21-SSIM (Mantiuk & Azimi, 2021; Wang et al., 2004), ΔE_{ITP} (ITU-R, 2019), HDR-VDP-3 (Mantiuk et al., 2023), HDR-LPIPS (Zhang et al., 2018) on PU21, and FR-HIDROVQA (Saini et al., 2024), a full-reference L_2 distance between two-scale normalized HIDRO-VQA features; the original regressor is unreleased.

Table 2: LumaEval frame win rates (%): fraction of aligned frames on which each method is best, with exact ties split equally. Bold and underline mark best and second best.

Method	PSNR	SSIM	ΔE_{ITP}	LPIPS	VDP-3	HIDRO
BT.2446c	14.2	<u>19.1</u>	16.5	7.2	<u>13.2</u>	6.5
Reinhard operator	<u>27.2</u>	**37.1**	**27.5**	<u>14.2</u>	11.2	3.5
HDRTVNet++	6.2	11.1	11.0	1.2	6.0	<u>17.0</u>
FMNet	13.2	13.1	14.0	0.2	1.2	3.0
ITM-LUT	3.2	0.1	3.0	7.2	2.2	1.0
VAE + RQS (control)	7.0	4.1	2.0	1.0	4.0	1.0
LUMAFLUX	**29.2**	15.1	<u>26.0</u>	**69.2**	**62.0**	**68.0**

On LumaEval, LUMAFLUX reaches 24.23 dB PU21-PSNR and 56.54 ΔE_{ITP}, the best result in every column of Table 1. The strongest analytic baseline trails by 0.49 dB and 5.06 color-error units, and the strongest CNN by 1.20 dB. A VAE+RQS control without diffusion transport collapses to 18.54 dB, showing that the learned transport, rather than output calibration alone, carries the gain. Appendices G–H break these results down by TMO family, degradation, and frame wins. Table 15 complements the macro-averages by measuring how often each method is best on individual aligned frames rather than only comparing dataset-level means.

The 50K-step HDRTV1K-specific checkpoint reaches 33.34 dB, 0.9427 SSIM, 0.9941 SR-SIM, 14.69 ΔE_{ITP}, and 7.962 HDR-VDP-3 on the 117-image test set under our PQ-code protocol. The mixed-corpus checkpoint reaches 27.52 dB (Appendix D).

On expert grades, LUMAFLUX reaches 33.31 dB and 0.544 FR-HIDROVQA. BT.2446c yields lower color error (37.83 versus 40.19). Native HDRTV1K exposes a distribution gap: the mixed-corpus model reaches 23.56 dB, above X2HDR's 18.65 dB (Appendix Tables 10 and 11).

Synthetic-track references contain 32.3% of pixels above the 100-nit SDR ceiling (Figure 8). BT.2446c under-expands to 19.1%, while the three CNNs over-expand to 47.9–48.6%. LUMAFLUX

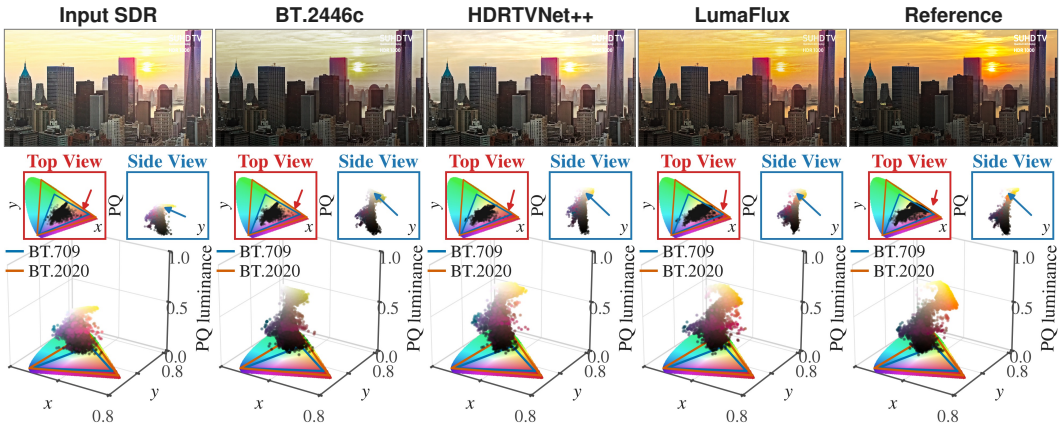

Figure 7: Rendered frames and CIE xyY volumes on an aligned city scene. Samples rise from the CIE 1931 diagram with BT.709 and BT.2020 triangles; height denotes PQ-coded luminance.

reaches 37.5% and gives the highest overlap with the reference highlight region (IoU 0.579; Reinhard 0.563, HDRTVNet++ 0.538). Reinhard is closer in overall share (29.7%), while the CNNs have lower error inside reference highlights (0.81–0.82 versus 0.87 stops).

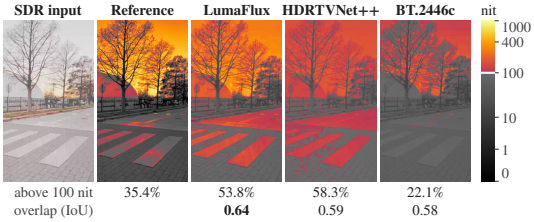

Figure 8: Headroom maps for one LumaEval frame; gray denotes 100 nit or less.

On 30 expert-track frames at 540p, 720p, and 1080p (Figure 1; one GH200, batch size 1; X2HDR and LEDiff full-frame), LUMAFLUX achieves the lowest FR-HIDROVQA at all three resolutions (0.431, 0.477, 0.567), respectively 22.9%, 21.2%, and 8.1% below the best measured baseline. At 2.64, 3.08, and 5.29 s per frame, it is 2.4–4.7× faster than X2HDR and 4.0–16.1× faster than LEDiff; HDRTVNet++ remains 173–201× faster.

We compress all 117 native frames with matched intra-only HEVC at QP 27, 32, 37, and 42 (Xu et al., 2023). Figure 9 shows that LUMAFLUX loses the least PQ-PSNR: from QP 27 to 42, the mixed-corpus model drops 0.55 dB (27.29 to 26.75) and the in-domain checkpoint 1.61 dB (33.94 to 32.33), versus 2.77 dB for HDRTVNet++ (35.90 to 33.13). Robustness is distinct from absolute quality: HDRTVNet++, trained in-domain, retains higher PSNR at every QP. Appendix Table 13 lists all values.

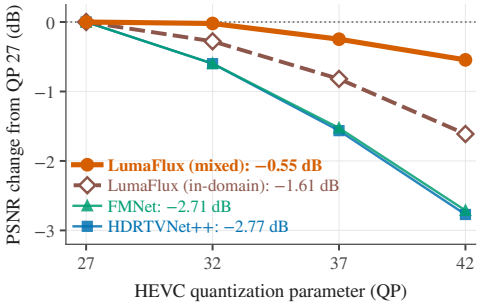

Figure 9: PQ-PSNR change from QP 27 on the 117 HDRTV1K pairs; flatter is more robust.

Figure 7 shows that the fixed operator and HDRTVNet++ under-fill the reference luminance-gamut volume, whereas LUMAFLUX recovers most of it. In Figure 10, LUMAFLUX has the highest whole-frame PU21-PSNR on all three examples: 30.83 dB on Nature (BT.2446c 30.70), 26.13 dB on Monument (HDRTVNet++ 25.68), and 28.22 dB on Sunset (BT.2446c 24.67). LEDiff (7.90 dB on Nature) and X2HDR (14.36 dB on Monument) are lowest. Appendix K adds scenes, error maps, and a ±4 EV exposure sweep.

Table 3 builds the design from rows (a)–(g), then modifies the output head in (h)–(i) or removes modulation in (j). Without RQS, LoRA-only (a) and the full model (i) both remain near 19 dB, confirming that the frozen VAE does not provide a calibrated HDR output. The monotone RQS (g) gains 9.69 dB over no RQS (i) and, relative to a linear spline (h), 0.45 dB and 4.35 color-error units. Spectral gating, (c)→(d), adds 0.98 dB, while removing timestep-layer modulation (j) costs 0.81 dB. The full design (g; 29.19 dB, 27.81 ΔE_{ITP}) is 0.83 dB above attention+MLP LoRA (b). Appendix C adds nine capacity, loss, and data-mix screens. On absolute temporal-energy deviation

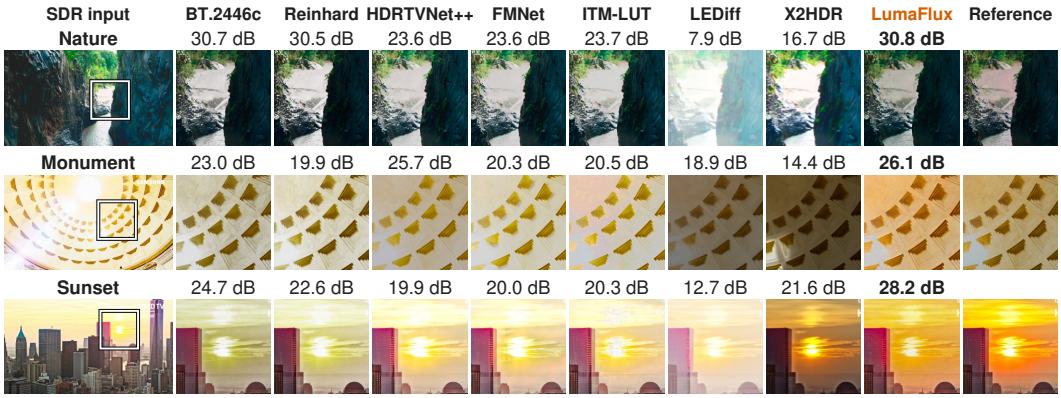

Figure 10: SDR inputs and corresponding method crops. Numbers give full-frame PU21-PSNR (dB), best in bold. Compare river shadow detail (Nature), coffer shading (Monument), and the sun disk and sky gradient (Sunset).

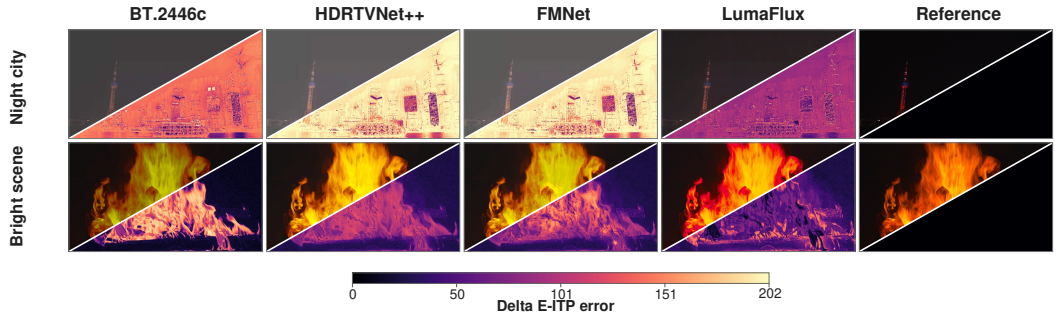

Figure 11: Each panel splits diagonally into the rendered preview and its per-pixel ΔE_{ITP} error on a shared scale. The maps localize color error around the waterfront and flame cores. LUMAFLUX reduces error in many selected regions, while its flame rendition retains visible chroma overshoot.

(Appendix Table 8, Figure 15), training-free stabilization makes LUMAFLUX comparable to the strongest framewise CNNs and closer to the reference than the analytic operator; pooled raw flicker remains within 3.27% of the reference.

7 CONCLUSION

We presented LUMAFLUX, a physically and perceptually guided adaptation of a frozen flow transformer for prompt-free conversion of 8-bit SDR to PQ BT.2020 HDR, together with the common LumaEval protocol. LUMAFLUX achieves the best result in every metric of the primary comparison, while ablations show that the monotone RQS decoder is critical for calibrated HDR reconstruction. Shared bridge noise and RQS-parameter smoothing extend the same image model to video, reducing signed excess flicker by 53.3% without temporal layers or video training. Because conditioning comes directly from the input and the generative prior remains frozen, the same 71.3M trainable parameters also support generate-then-lift text-to-HDR (Appendix E) and efficient specialization by retraining only the adapters. Limitations remain: in-domain CNNs outperform the mixed-corpus model on native HDRTV1K pairs, and eight-step sampling remains slower than single-pass baselines (Appendix N).

AI USE STATEMENT

Generative AI assistants were used during preparation of this submission for format conversion, manuscript editing, and code to visualize stored outputs. This assistance did not produce new model outputs or experimental measurements. The research and experiments preceded this submission-preparation use. The authors are responsible for the final content and interpretation of this work.

ETHICS STATEMENT

Inverse tone mapping estimates information absent from an SDR observation. Generated highlight detail should therefore be treated as a reconstruction, especially in archival, documentary, or evidentiary settings. Restricted source media are represented by content identifiers, curation scripts, and checksums rather than redistributed without permission. Appendix N discusses hallucination, display dependence, and limits of the evaluation.

REPRODUCIBILITY STATEMENT

Sections 4–6 specify the model, training corpus, and shared evaluation conventions. Appendices A, C, and L provide adapter details, the complete ablation matrix, and metric and baseline protocols. Appendix M gives the inference algorithm. Code will released.

REFERENCES

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Bowen Chen, Shreshth Saini, Balu Adsumilli, and Alan C. Bovik. LumaGuide: Distribution shaping for training-free HDR generation in diffusion models. *arXiv preprint arXiv:2607.26237*, 2026.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. HDRUNet: Single image HDR reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 354–363, 2021a.

Xiangyu Chen, Zhengwen Zhang, Jimmy S. J. Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from sdrtv to hdrtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4500–4509, 2021b.

Xiangyu Chen, Zheyuan Li, Zhengwen Zhang, Jimmy S. Ren, Yihao Liu, Jingwen He, Yu Qiao, Jiantao Zhou, and Chao Dong. Towards efficient sdrtv-to-hdrtv by learning from image formation. *arXiv preprint arXiv:2309.04084*, 2023. URL https://arxiv.org/abs/2309.04084.

CTA. Television technology consumer definitions. [Online], 2024. Accessed February 2024.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6), 2017.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Cheng Guo, Leidong Fan, Ziyu Xue, and Xiuhua Jiang. Learning a practical sdr-to-hdrtv up-conversion using new dataset and degradation models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22231–22241, 2023a. doi: 10.1109/CVPR52729.2023.02129.

Cheng Guo, Leidong Fan, Qian Zhang, Hanyuan Liu, Kanglin Liu, and Xiuhua Jiang. Redistributing the precision and content in 3D-LUT-based inverse tone-mapping for HDR/WCG display. In *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production*, pp. 1–10, 2023b. doi: 10.1145/3626495.3626503.

Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai. Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 2890–2898, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548043. URL https://doi.org/10.1145/3503161.3548043.

Gang He, Kepeng Xu, Li Xu, Wenxin Yu, and Xianyun Wu. Beyond feature mapping gap: Integrating real HDRTV priors for superior SDRTV-to-HDRTV conversion. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1089–1097, 2025. doi: 10.24963/ijcai.2025/122.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Peihuan Huang, Gaofeng Cao, Fei Zhou, and Guoping Qiu. Video inverse tone mapping network with luma and chroma mapping. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 1383–1391, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612199. URL https://doi.org/10.1145/3581783.3612199.

ITU-R. Bt.2020-2: Parameter values for ultra-high definition television systems for production and international programme exchange. Technical Report BT.2020-2, International Telecommunication Union, October 2015a. URL https://www.itu.int/rec/r-rec-bt.2020.

ITU-R. Bt.709-6: Parameter values for the hdtv standards for production and international programme exchange. Technical Report BT.709-6, International Telecommunication Union, June 2015b. URL https://www.itu.int/rec/R-REC-BT.709-6-201506-I/en.

ITU-R. Bt.2407-0: Colour gamut conversion from recommendation itu-r bt.2020 to recommendation itu-r bt.709. Technical Report BT.2407-0, International Telecommunication Union, October 2017.

ITU-R. Bt.2100-2: Image parameter values for high dynamic range television for use in production and international programme exchange. Technical Report BT.2100-2, International Telecommunication Union, July 2018. URL https://www.itu.int/rec/r-rec-bt.2100.

ITU-R. Bt.2124-0: Objective metric for the assessment of the potential visibility of colour differences in television. Technical Report BT.2124-0, International Telecommunication Union, January 2019. URL https://www.itu.int/.

ITU-R. Bt.2446-1: Methods for conversion of high dynamic range content to standard dynamic range content and vice-versa. Technical Report BT.2446-1, International Telecommunication Union, March 2021.

Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3116–3125, 2019.

Soo Ye Kim, Jihyong Oh, and Munchurl Kim. JSI-GAN: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11287–11295, 2020.

Naomi Ken Korem, Mohamed Oumoumad, Harel Cain, Matan Ben Yosef, Urska Jelercic, Ofir Bibi, Yaron Inger, Or Patashnik, and Daniel Cohen-Or. HDR video generation via latent alignment with logarithmic encoding. *arXiv preprint arXiv:2604.11788*, 2026.

Xianwei Li, Weimin Zhang, Huiyuan Fu, and Huadong Ma. Diffusion-based sdrtv-to-hdrtv reconstruction with artifact suppression. SSRN Electronic Journal, March 2024. URL https://ssrn.com/abstract=5549060.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Rafał K Mantiuk and Maryam Azimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*, pp. 1–5. IEEE, 2021.

Rafał K Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Netflix. Netflix open content. Creative Commons Attribution 4.0 test content, 2026. URL https://opencontent.netflix.com/home. Accessed 2026-08-05.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023.

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, July 2002. ISSN 0730-0301. doi: 10.1145/566654.566575. URL https://doi.org/10.1145/566654.566575.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. RB-Modulation: Training-free stylization using reference-based modulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bnINPG5A32.

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*, pp. 469–479, 2024.

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)*, pp. 2504–2509. IEEE, 2025.

Shreshth Saini, Bowen Chen, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. Seeing beyond 8bits: Subjective and objective quality assessment of HDR-UGC videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15538–15549, June 2026a.

Shreshth Saini, Bowen Chen, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. BrightRate: Quality assessment for user-generated HDR videos. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1522–1532, 2026b.

Tong Shao, Deming Zhai, Junjun Jiang, and Xianming Liu. Hybrid conditional deep inverse tone mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

SMPTE. St 2084:2014 - high dynamic range electro-optical transfer function of mastering reference displays. Technical Report ST 2084:2014, Society of Motion Picture and Television Engineers (SMPTE), 2014.

Chinmay Talegaonkar, Jinshi He, Christopher McKenna, and Nicholas Antipa. Single-shot HDR recovery via a video diffusion prior, 2026. URL https://arxiv.org/abs/2605.11628.

SaiKiran Tedla, Francesco Banterle, Trevor Canham, Karanpreet Raja, David B. Lindell, Kiriakos N. Kutulakos, Jiacheng Li, Feiran Li, and Daisuke Iso. Generating HDR video from SDR video, 2026. URL https://arxiv.org/abs/2605.14703.

Abhinau K Venkataramanan and Alan C Bovik. Subjective quality assessment of compressed tone-mapped high dynamic range videos. *IEEE Transactions on Image Processing*, 2024.

Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Hans-Peter Seidel, Christian Theobalt, Karol Myszkowski, and Thomas Leimkühler. GlowGAN: Unsupervised learning of HDR images from LDR images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Chao Wang, Francesco Banterle, Bin Ren, Radu Timofte, Xin Lu, Yufeng Peng, Chengjie Ge, Zhijing Sun, Ziang Zhou, Zihao Li, et al. AIM 2025 challenge on inverse tone mapping report: Methods and results, 2025a. URL https://arxiv.org/abs/2508.13479.

Chao Wang, Zhihao Xia, Thomas Leimkühler, Karol Myszkowski, and Xuaner Zhang. Lediff: Latent exposure diffusion for hdr generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 453–464, 2025b.

Hu Wang, Mao Ye, Xiatian Zhu, Shuai Li, Ce Zhu, and Xue Li. KUNet: Imaging knowledge-inspired single HDR image reconstruction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 1408–1414, 2022. doi: 10.24963/ijcai.2022/196.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Ronghuan Wu, Wanchao Su, Kede Ma, Jing Liao, and Rafał K. Mantiuk. X2HDR: HDR image generation in a perceptually uniform space, 2026. URL https://arxiv.org/abs/2602.04814.

Gang Xu, Qibin Hou, Le Zhang, and Ming-Ming Cheng. FMNet: Frequency-aware modulation network for SDR-to-HDR translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Kepeng Xu, Li Xu, Gang He, Xianyun Wu, Zhiqiang Zhang, Wenxin Yu, and Yunsong Li. Dual inverse degradation network for real-world sdrtv-to-hdrtv conversion, 2023. URL https://arxiv.org/abs/2307.03394.

Li Xu, Siqi Wang, Kepeng Xu, Lin Zhang, Gang He, Weiran Wang, and Yu-Wing Tai. RealRep: Generalized SDR-to-HDR conversion via attribute-disentangled representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11305–11313, 2026. doi: 10.1609/aaai.v40i13.38111. URL https://ojs.aaai.org/index.php/AAAI/article/view/38111.

Zhengming Yu, Li Ma, Mingming He, Leo Isikdogan, Yuancheng Xu, Dmitriy Smirnov, Pablo Salamanca, Dao Mi, Pablo Delgado, Ning Yu, Julien Philip, Xin Li, Wenping Wang, and Paul Debevec. DiffHDR: Re-exposing LDR videos with video diffusion models, 2026. URL https://arxiv.org/abs/2604.06161.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

APPENDIX

A Additional Method Details . . . 15
B Measured Relative-Structure Bias . . . 18
C Complete Ablation Study . . . 18
D Additional Benchmark Tables . . . 19
E Text-to-HDR Video via Generate-Then-Lift . . . 21
F Track-Specific and Compression Results . . . 22
G Luma-Eval by TMO Family . . . 24
H Luma-Eval Frame Win Rates . . . 25
I Per-Variant Luma-Eval Results . . . 25
J Efficiency and Model Scale . . . 26
K Additional Qualitative Diagnostics . . . 27
L Evaluation Protocol Details . . . 28
M Inference Procedure . . . 29
N Limitations and Broader Impact . . . 30
O Corpus Construction and Signal Normalization . . . 31
P Supplementary Related Work . . . 32

A ADDITIONAL METHOD DETAILS

A.1 COLOR PROCESSING

The SDR formation model in Equation 1 compresses both dynamic range and gamut. Inverse tone mapping must invert the BT.709 transfer function, map BT.709 primaries into BT.2020, and expand display-referred tone without introducing banding or chromatic discontinuities. All metric implementations consume explicit color metadata, and perceptual comparisons are performed after conversion into their required luminance or color spaces.

A.2 ADAPTER PLACEMENT AND SHAPES

The prompt-free implementation uses learned null context tokens concatenated with projected SigLIP tokens, plus a learned pooled vector conditioned by physical statistics. The Null and Pooled labels in Figures 2 and 3 denote this conditioning interface. The physical path decodes SDR to linear light, converts primaries to BT.2020, and then extracts luminance and descriptors; r denotes pooled Fourier bands and g the global-statistics embedding.

PGA applies the low-rank attention update in Equation 6: the physical gate has one value per token and head, the spectral gate one value per head, and each head's value is repeated over its channels. PCM predicts its modulation coefficients through Equation 7 and adds the normalized, modulated branch to the incoming hidden state as in Equation 8. PCM is applied before each frozen block, PGA acts within its attention projections, and the HDR residual coupler follows the block. The physical projection, perceptual connector, and residual coupler use bottleneck width 64. Zero initialization of the terminal PCM and coupler projections makes the initial adapter path an identity perturbation.

A.3 TONE-FIELD INTERPRETATION

Let $Y_{\text{out}} \in [0, 1]$ denote the decoded luma surrogate and let $g_{\boldsymbol{\theta}}$ be a monotone RQS on the same interval. Piecewise rational splines approximate a smooth monotone tone curve while keeping derivatives positive. This structure does not recover detail absent from the SDR input, but it is an interpretable, differentiable mechanism for calibrating the VAE output to HDR luminance.

A.4 VAE ROUND-TRIP INPUT ENCODINGS

The frozen FLUX VAE was never trained on display-referred HDR, so the choice of input encoding determines how much of the PQ signal survives an encode-decode round trip. A controlled selection study passes six registered LIVE-TMHDR crops through the frozen VAE twice, once with PQ-coded input and once with PU21-coded input, and scores each reconstruction against its source. Across the six frames, the PQ encoding averages 39.55 dB PU21-PSNR and 7.91 ΔE_{ITP}, and the PU21 encoding averages 39.19 dB and 7.56. Figure 12 shows all six frames: PQ wins PU21-PSNR on

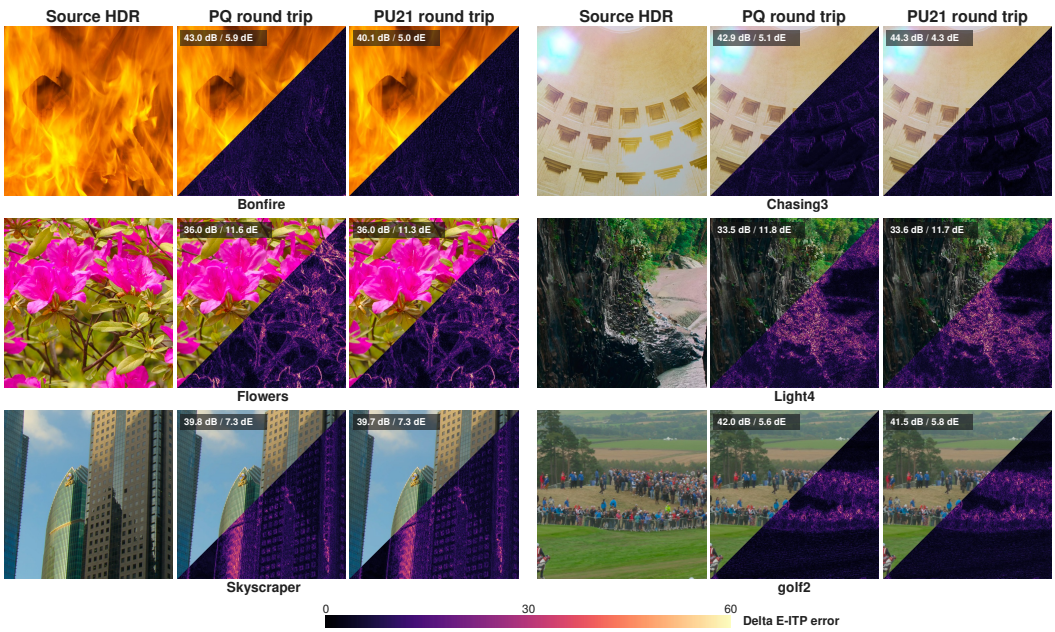

Figure 12: Frozen-VAE round trips on the six registered LIVE-TMHDR crops. Each group shows the source, then the PQ and PU21 reconstructions split diagonally with their per-pixel ΔE_{ITP} error; chips give the registered PU21-PSNR and mean error. PQ wins Bonfire by 2.90 dB and PU21 wins Chasing3 by 1.33 dB; the other four frames differ by at most 0.6 dB.

four of six and on the mean, PU21 wins mean color error by 0.35 ΔE_{ITP}, and the pipeline retains the PQ path under the selection rule of higher mean PU21-PSNR with color error as the tie-breaker. Round-trip error concentrates in high-frequency texture rather than in smooth tone regions, and even the hardest frame retains 33.5 dB. Precision does not add artifacts at this stage: on a smooth-gradient video frame, bf16 and fp32 VAE round trips score 41.81 and 41.89 dB against the source and differ from each other by 57 dB, with no visible banding under contrast stretching, unlike the video-VAE precision artifacts reported by Yu et al. (2026). This bounded, texture-dominated error motivates the design split in which the diffusion transport carries structure while the monotone RQS head calibrates tone: the spline corrects the global luminance placement that the VAE alone cannot guarantee, without asking the VAE to represent out-of-distribution amplitudes.

A follow-up study repeats the round trip under four input encodings at evaluation-set level, over the 100 registered Luma-Eval synthetic-track reference frames with centered 1024^2 crops; the analysis style follows Korem et al. (2026). Figure 13 compares PQ, PU21, HLG, and linear 1,000-nit-normalized coding of the identical sources. Linear coding collapses to 30.85 dB and 27.01 ΔE_{ITP} because most scene content is compressed against zero, which also produces by far the largest divergence from the SDR-prior distribution in both pixel and latent space. The three perceptual codings stay within 0.44 dB of one another: PQ and HLG tie on PSNR while HLG sits closest to the SDR prior at this level. The prior pools the paired degraded SDR inputs of the same frames, and the pipeline retains the PQ path fixed by the earlier six-crop se-

Table 3: Component ablations (20k steps, 55-video pool, same 82 held-out frames). (a) LoRA only; (b) attention and MLP LoRA; *lin.*: linear spline. Superscripts: change from row (g); † marks two seed-0 runs of the full configuration. The RQS head has 21.9K parameters and Ψ 41.1K.

	PGA	Gate	PCM	Coup.	T-mod	RQS	PSNR↑	ΔE_{ITP} ↓
(a)	×	×	×	×	✓	×	$19.29^{\downarrow 9.90}$	$256.84^{\uparrow 229.03}$
(b)	×	×	×	×	✓	✓	$28.36^{\downarrow 0.83}$	$36.35^{\uparrow 8.53}$
(c)	✓	×	×	×	✓	✓	$28.57^{\downarrow 0.62}$	$33.48^{\uparrow 5.66}$
(d)	✓	✓	×	×	✓	✓	$\mathbf{29.56}^{\uparrow 0.37}$	$\underline{28.38}^{\uparrow 0.57}$
(e)	✓	✓	✓	×	✓	✓	$\underline{29.20}^{\uparrow 0.01}$	$31.76^{\uparrow 3.95}$
(f)†	✓	✓	✓	✓	✓	✓	$28.67^{\downarrow 0.52}$	$35.00^{\uparrow 7.18}$
(g)†	✓	✓	✓	✓	✓	✓	$29.19^{(\mathrm{ref.})}$	$\mathbf{27.81}^{(\mathrm{ref.})}$
(h)	✓	✓	✓	✓	✓	*lin.*	$28.74^{\downarrow 0.45}$	$32.16^{\uparrow 4.35}$
(i)	✓	✓	✓	✓	✓	×	$19.50^{\downarrow 9.69}$	$261.78^{\uparrow 233.97}$
(j)	✓	✓	✓	✓	×	✓	$28.38^{\downarrow 0.81}$	$33.90^{\uparrow 6.09}$

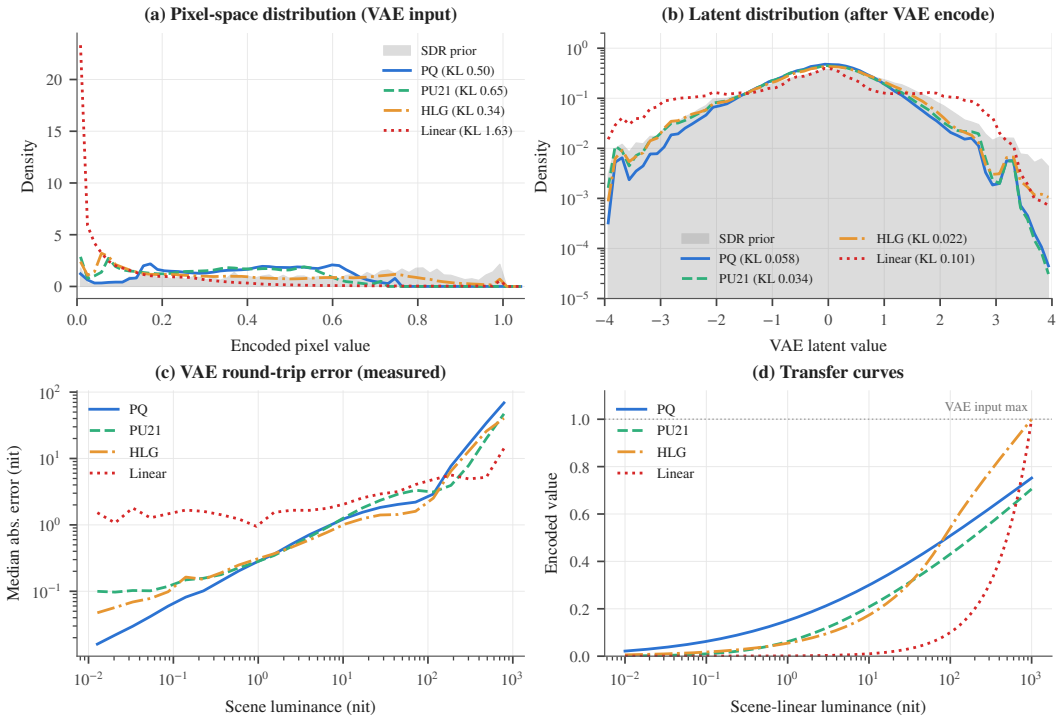

Figure 13: Input encodings for the frozen VAE over the 100 registered Luma-Eval reference frames, in the style of Korem et al. (2026). (a) Pixel and (b) latent distributions against the pooled SDR prior, with KL divergences in the legends; (c) median round-trip luminance error in log-luminance bins; (d) the transfer curves. Linear coding starves mid-tones and collapses, while the three perceptual codings stay close to the prior the frozen VAE expects.

lection rule. An archived rerun of the six-
crop subset agrees with the earlier measurements within 0.14 dB.

We normalize decoded and reference luma by the PQ code of the mastering peak. Let $U \in [0,1]$ be an ideal decoded luma, $Y^\star = g^\star(U)$ its reference, and $X = U + \varepsilon \in [0,1]$ the observed decoded luma, with $\mathbb{E}[\varepsilon^2] \le \sigma^2$. Assume that $g^\star$ is strictly increasing and M-Lipschitz on $[0,1]$, with fixed endpoints 0 and 1. This isolates a global tone distortion for fixed scene conditioning; spatial reconstruction remains the transport model's responsibility.

Proposition (calibration error under decoded-luma perturbations). Let $\mathcal{G}_K$ be a class of monotone rational-quadratic splines with K bins, positive knot derivatives, and fixed endpoints (Durkan et al., 2019). Suppose it contains $\widetilde{g}$ satisfying $\|\widetilde{g} - g^\star\|_\infty \le \epsilon$. If $\widehat{g}$ minimizes the population risk $\mathcal{R}(g) = \mathbb{E}[(g(X) - Y^\star)^2]$ over $\mathcal{G}_K$, then, writing μ for the distribution of X,

$$\sqrt{\mathcal{R}(\widehat{g})} \le \epsilon + M\sigma, \qquad \|\widehat{g} - g^\star\|_{L^2(\mu)} \le \epsilon + 2M\sigma. \tag{14}$$

In particular, when $\sigma = 0$ and $g^\star \in \mathcal{G}_K$, every population minimizer equals $g^\star$ μ-almost everywhere.

Proof sketch. The Lipschitz assumption gives $|g^\star(X) - Y^\star| = |g^\star(U+\varepsilon) - g^\star(U)| \le M|\varepsilon|$. The triangle inequality in L^2 therefore yields

$$\sqrt{\mathcal{R}(\widehat{g})} \le \sqrt{\mathcal{R}(\widetilde{g})} \le \|\widetilde{g}(X) - g^\star(X)\|_2 + \|g^\star(X) - Y^\star\|_2 \le \epsilon + M\sigma,$$
$$\|\widehat{g}(X) - g^\star(X)\|_2 \le \|\widehat{g}(X) - Y^\star\|_2 + \|Y^\star - g^\star(X)\|_2 \le \epsilon + 2M\sigma.$$

For the noiseless, representable case, the first inequality gives zero risk, so the functions agree on the observed support. Parameters and values outside that support need not be unique. □

The approximation error ϵ is an explicit assumption for the chosen spline class; its value depends on the tone function and bin budget. The bound separates this error from decoded-luma perturbations and motivates a dedicated calibration head. The implementation enforces positive bin sizes

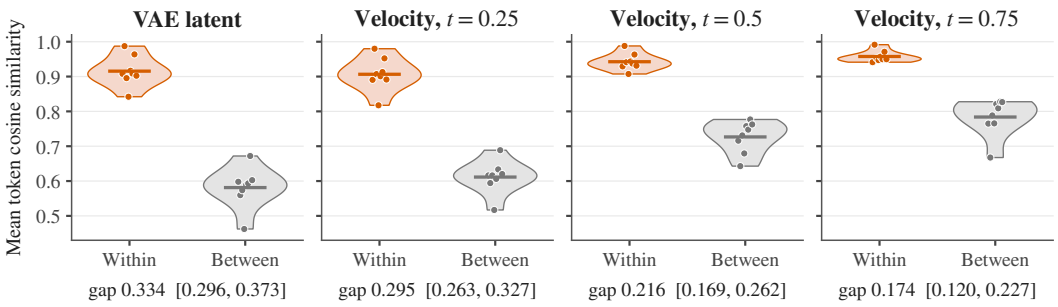

Figure 14: Within-content tone-variant similarity against the between-content control for frozen-VAE tokens and for velocity directions at three timesteps. Dots are the eight frame-level means and bars mark group means; the text under each panel gives the paired within-minus-between gap with its 95% interval. The gap is positive in all four panels.

and slopes; M bounds the assumed reference distortion, and learned slopes have no upper cap. Full reconstruction also depends on the spatial transport and the combined training objective in Equation 12, which includes linear-light L_1 terms rather than this illustrative squared-luma risk. The empirical structure measurement in Appendix B and the tone-head ablation provide separate evidence for those parts of the design.

B Measured Relative-Structure Bias

We test the motivating claim that a frozen generative prior preserves content structure across monotone tone transformations rather than assuming it from architecture alone. The experiment applies six deterministic PQ-rendered tone variants to eight fixed LIVE-TMHDR crops. No LumaFlux adapter or SigLIP conditioning is constructed. Every content, tone variant, and timestep shares one noise tensor, which removes noise realization as an alternative explanation.

For each content, we compute cosine similarity between packed spatial tokens for all pairs of tone variants. This is the *within-content* distribution. The control compares different contents under the same tone variant, giving the *between-content* distribution. We report the same contrast for frozen-VAE latents and for raw FLUX velocity directions at $t \in \{0.25, 0.5, 0.75\}$. Similarity is computed per token and then averaged within a frame. Confidence intervals are two-sided 95% Student t intervals over the eight frame-level within-minus-between gaps.

Figure 14 shows that velocity directions are more similar across tone mappings of the same content than across different contents at every tested timestep. The paired confidence interval for the within-minus-between gap remains strictly positive at $t = 0.25$, 0.5, and 0.75. The latent comparison exhibits the same ordering. This measurement supports a relative-structure bias in the frozen prior under the tested transformations, but it does not imply invariance to arbitrary tone operators or guarantee recovery of clipped HDR detail.

C Complete Ablation Study

The ablation study was conducted on 55 videos (approximately 2.6% of the 2,092-video training corpus) at reduced steps, with the full evaluation protocol. The component matrix in the main paper (Table 3) and Tables 4 and 5 here report all 19 completed configurations; the data-mix table repeats the full-design reference. The UGC-only and PGC-only rows use subsets of this 55-video pool. The two full-configuration runs share seed 0, so their gap is run-to-run variation rather than a multi-seed estimate. The 20,000-step A and B rows study components and alternative designs; the 15,000-step C rows vary capacity, loss terms, and data mix. All rows use the common 82-frame holdout, but comparisons across the two budgets are not matched-compute estimates.

The component matrix of Table 3 marks PGA, the spectral gate, PCM, the HDR coupler, timestep-layer modulation, and the RQS tone head; "lin." denotes a linear spline. The two † rows are separate runs of the full configuration, both with seed 0; their 0.52 dB gap measures run-to-run variation at this budget. The output mapping dominates every other choice: LoRA only, row (a), and the full

Table 4: Seven capacity and loss screens at 15,000 steps on the same 82-frame holdout. Marks rank these seven rows only.

Variant (15k steps)	PSNR↑	ΔE_{ITP} ↓
Adapter capacity		
Bridge noise 0.02	28.53	31.58
Adapter rank 4	28.34	33.55
Adapter rank 16	28.44	32.07
Bottleneck width 32	**29.45**	**26.55**
Bottleneck width 128	28.69	31.14
Losses		
Velocity-only loss	28.01	37.47
No spline-smooth loss	28.47	32.88

model without the RQS head, row (i), stay near 19 dB, while every row with a spline head lies between 28.36 and 29.56 dB. The capacity and loss screens follow separately because their training budget differs.

The capacity and loss screens of Table 4 use 15,000 reduced steps. The velocity-only screen trains with $\mathcal{L}_{\mathrm{FM}}$ alone; the no-spline-smooth screen removes only the λ_3 term. Bottleneck width 32 gives the strongest color result in this budget, 26.55 ΔE_{ITP}; width 128 raises color error to 31.14. Rank 4 and rank 16 are close at 28.34 and 28.44 dB. Removing the spline-smoothness term costs 0.81 ΔE_{ITP} relative to the rank-16 configuration, and velocity-only training reaches 28.01 dB and 37.47 color error. These rows are not matched to the 20,000-step component model.

We train with three corpus compositions: UGC only, PGC only, and the registered mix of both halves (Table 5). UGC-only training reaches 29.21 dB and 27.09 ΔE_{ITP} on the UGC-weighted holdout, PGC-only drops to 23.02 dB and 56.05, and the combined corpus reaches 29.19 dB and 27.81 while retaining professional-content coverage. The 6.19 dB UGC-versus-PGC difference describes this holdout, not general data efficiency. We keep the mixed corpus because the synthetic, expert-grade, and native-pair tracks require cross-source coverage.

Table 5: Training data mixes under the ablation protocol, scored on the UGC-weighted 82-frame holdout. The shaded row is the corpus used for the final model.

Training corpus	PSNR↑	ΔE_{ITP} ↓
UGC only	**29.21**	**27.09**
PGC only	23.02	56.05
UGC + PGC	29.19	27.81

D Additional Benchmark Tables

This section collects the decoded-dataset statistics, the HDRTV1K comparison with published numbers, and the temporal evidence referenced by the main text.

Table 6: HDR dataset statistics computed on decoded PQ/BT.2020 frames. FHLP and EHL measure highlight proportion and extent, FWGP and EWG wide-gamut proportion and extent (Guo et al., 2023a); these four are percentages. SI is spatial information, stdL the luminance spread, Colorful. the colorfulness, and N_s the number of decoded frames. Bold marks the largest value per statistic.

Dataset	FHLP↑	EHL↑	FWGP↑	EWG↑	SI↑	stdL↑	Colorful.↑	N_s	Pairs	Format (PQ/BT.2020)
HDRTV1K test	24.880	11.170	**26.336**	**1.482**	33.646	11.757	**27.411**	117	117	16-bit PNG
LIVE-TMHDR expert track	27.811	12.410	15.087	1.289	37.299	12.302	25.351	100	100	16-bit uint tensor
Luma-Eval corpus	**31.954**	**13.094**	13.950	1.019	**39.319**	**13.078**	22.502	2,000	314,396	16-bit uint tensor

Table 6 shows that the three sources are complementary. The Luma-Eval corpus has the largest highlight proportion and extent, spatial information, and luminance spread, while the HDRTV1K test set uses the wide gamut most and is the most colorful. The LIVE-TMHDR expert track lies between the two on every statistic.

Section L gives the per-row provenance of Table 7. The HDRTV1K-specific LumaFlux checkpoint stays below the best published PSNR (33.34 against 38.60 dB), in line with the native-track result of the main text: networks trained on this distribution stay ahead. Track-specific reruns appear in Tables 10 and 11 because their metric suite and calibration differ from these literature values.

Table 7: HDRTV1K baseline values are quoted as reported and retain each author's evaluation protocol; the three HDRTVNet variants share one group row, and "–" means not reported. The LUMAFLUX row is the HDRTV1K-specific checkpoint under our 117-image protocol. Bold and underline mark the best and second-best value per column.

Method	PSNR↑	SSIM↑	SR-SIM↑	ΔE_{ITP} ↓	HDR-VDP-3↑
HDRTVNet family (Chen et al., 2021b; 2023)					
AGCM+LE+HG	37.21	0.9699	0.9968	9.11	8.569
AGCM-LE-HR++	38.36	0.9735	**0.9975**	8.28	<u>8.751</u>
AGCM-LE++, joint	**38.60**	0.9745	0.9973	**7.67**	8.696
HDCFM (He et al., 2022)	<u>38.42</u>	0.9732	<u>0.9974</u>	7.83	8.572
FMNet (Xu et al., 2022)	37.94	0.9747	0.9957	8.10	8.510
KUNet (Wang et al., 2022)	37.78	**0.9868**	0.9971	<u>7.80</u>	8.393
Deep SR-ITM (Kim et al., 2019)	37.10	0.9686	0.9950	9.24	8.233
ITM-LUT (Guo et al., 2023b)	36.69	<u>0.9811</u>	–	10.19	**8.912**
Beyond-FMG (He et al., 2025)	35.06	0.9609	–	–	–
RealRep (Xu et al., 2026)	29.12	0.9412	–	–	–
LUMAFLUX (in-domain)	33.34	0.9427	0.9941	14.69	7.962

Training-free stabilization

Sharing one bridge-noise realization across frames and smoothing the applied RQS parameters with an EMA turns the image model into a video converter with no temporal layers, optical flow, or video fine-tuning, cutting signed excess flicker by 53.3%.

Table 8 counts both failure directions as error: added flicker, and a low raw energy obtained by suppressing real temporal variation. Stabilization lowers the LUMAFLUX mean deviation from 1.344 to 1.099, between HDRTVNet++ (1.012) and FMNet (1.166) and well under the analytic operator (1.972). Figure 15 shows the corresponding luma traces. The pooled signed measure of the main text is separate: over the five clips, excess flicker energy falls from 547.4 to 255.5 (53.3%), and raw energy (8,075.5) stays within 3.27% of the reference (7,819.9), in units of 10^3.

Table 8: Temporal-energy deviation, $|\mathrm{flicker}_{\mathrm{pred}} - \mathrm{flicker}_{\mathrm{ref}}|$ in 10^6 PU21-energy units on five clips (02c3 to 0237 are CHUG content IDs). Smaller is closer to the reference; the gray reference row is zero by definition and is not ranked. Bold marks the closest method per column.

Method	Bonfire	02c3	055e	0392	0237	**Mean**
Reference	0.000	0.000	0.000	0.000	0.000	0.000
BT.2446c	7.813	1.411	0.450	**0.063**	**0.124**	1.972
HDRTVNet++	3.005	0.207	0.147	0.442	1.257	**1.012**
FMNet	3.853	**0.202**	**0.074**	0.579	1.124	1.166
LUMAFLUX	**1.992**	0.691	2.500	0.968	0.569	1.344
+ stabilization	2.109	0.469	1.676	0.879	0.362	1.099

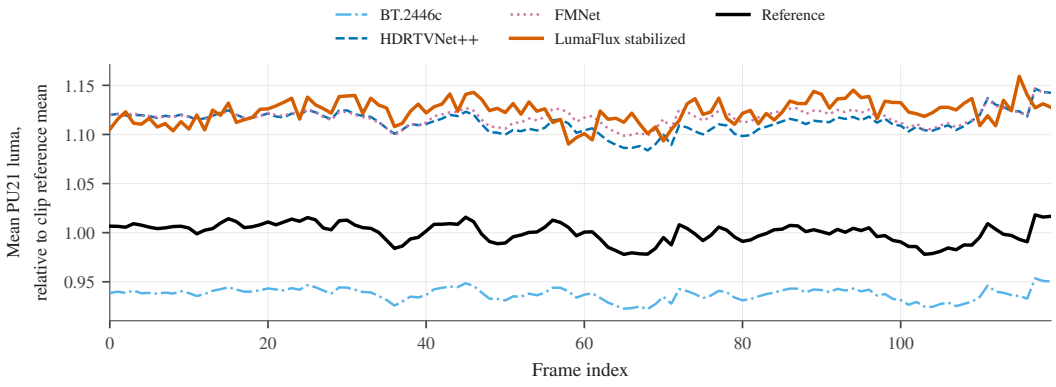

Figure 15: Per-frame mean PU21 luma averaged over the four complete clips (first 120 frames, each normalized by its reference mean). The unstabilized LUMAFLUX trace coincides with the stabilized one at this scale because stabilization acts on frame-to-frame differences, which Table 8 quantifies per clip.

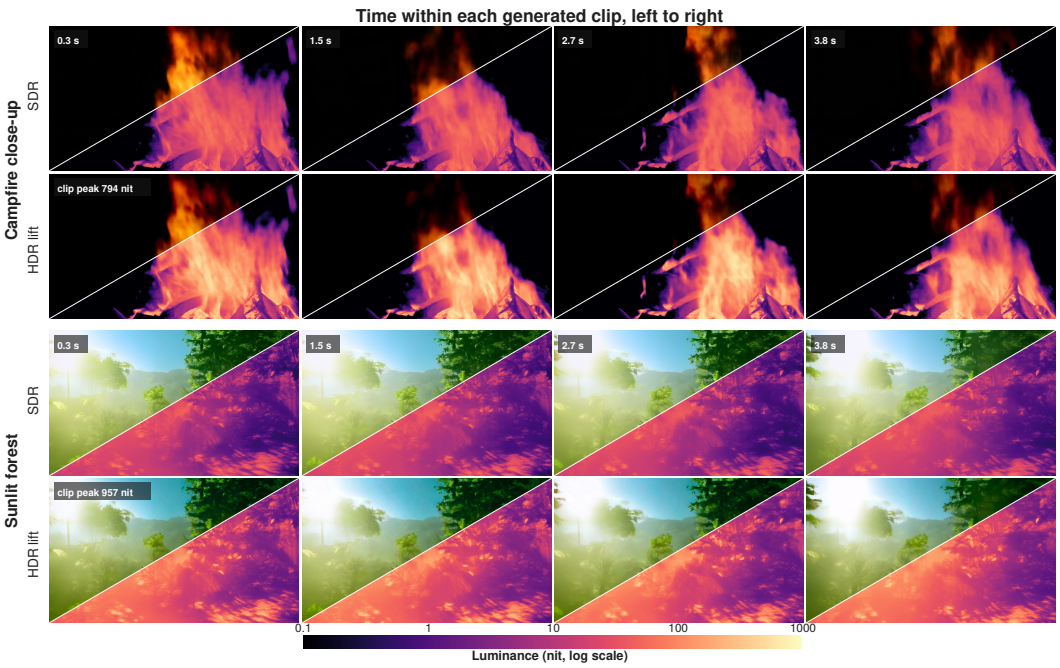

Figure 16: Generate-then-lift sequences at four time points; each panel splits diagonally into the rendered frame and its log-luminance map on the shared 0.1 to 1,000 nit scale. The SDR maps saturate at the 100-nit container ceiling, while the stabilized lift reaches the measured clip peaks (chips, from Table 9) in the flame cores and sun bloom.

E TEXT-TO-HDR VIDEO VIA GENERATE-THEN-LIFT

Any text-to-video generator becomes a text-to-HDR-video generator by lifting its SDR output with LUMAFLUX stabilized inference. We instantiate this composition with LTX-Video (`Lightricks/LTX-Video`): each 4.03-second clip contains 121 frames at 1216×704 and 30 fps, generated with 50 inference steps and guidance scale 3.0. Because no paired HDR reference exists for generated content, evaluation is intrinsic and measures output range, gamut expansion, source-relative temporal energy, and lift latency; Figure 16 shows lifted sequences, Table 9 the per-clip statistics, and Figure 17 where five of the lifts exceed the 100-nit SDR ceiling.

Table 9: Text-to-HDR lift statistics for stabilized eight-step LUMAFLUX lifts of six LTX-Video SDR clips. DR is $\log_2(L_{99.5}/L_{0.5})$, peak is the 99.99th luminance percentile, and the gamut column is the pixel fraction outside BT.709 after mapping the linear BT.2020 output. Flicker is the mean transition-wise L1 PU21 luminance energy in 10^6 units, and Ratio is lift over source. Lift time is 5.82 s per frame on every clip (5.817 to 5.830).

	DR	Peak	Outside	**PU21 flicker energy**		
Clip	(stops)	(nit)	709 (%)	Source	Lift	Ratio
Ocean sunset	9.59	955.7	0.79	5.724	8.057	1.407
Night fireworks	16.39	955.6	6.70	2.142	3.131	1.461
Neon rain street	13.08	957.1	11.71	7.309	11.082	1.516
Campfire close-up	16.47	793.6	46.52	12.151	18.606	1.531
Sunlit forest	9.40	957.3	42.82	2.277	3.928	1.725
Blue-hour skyline	10.60	739.0	86.27	1.270	1.940	1.528

In Table 9 every lift peaks between 739.0 and 957.3 nit, 7.4 to 9.6 times the 100-nit SDR ceiling. Lift and source flicker energies have Pearson correlation 0.998: their ratio stays between 1.407 and 1.725 while source flicker spans 1.270 to 12.151, so the stabilized lift tracks source-driven temporal variation rather than replacing its scene ordering. Latency counts LUMAFLUX inference only; generation, encoding, and metrics are excluded.

Figure 17 shows where the lift spends its headroom for five of the six clips, each at the stored keyframe (frame 0, 60, or 120 of 121) with the largest area above the 100-nit SDR ceiling; the blue-hour skyline, where that area stays below 0.5% of the frame, is omitted. In the shown frames 5.4%, 11.6%, 14.8%, 33.5%, and 42.1% of pixels exceed the ceiling. Pixels above 400 nit cover 1.4% to 4.6% of the frame in four clips and 12.4% in the sunlit forest, so outside that scene the peaks in Table 9 are carried by a small emitter area: firework bursts, flame cores, signs and their reflections, the solar disk. Over the 18 stored keyframes of all six clips the share above the ceiling ranges from 0.4% to 42.1%, and 83.9% to 98.4% of those pixels belong to the equally large set of brightest source pixels: the lift places highlights on the source's own light sources rather than on new regions.

F Track-Specific and Compression Results

This section supplies the complete expert-grade, native-pair, pooled-CRF, and fixed-QP evidence referenced in Section 6.

Table 10: Seven metrics on the expert-grade track: the 100 LIVE-TMHDR pairs whose SDR side is an expert grade. Core columns use the preregistered per-content exposure calibration; the optional metrics use the shared raw-output implementation. Bold and underline mark the best and second-best value per column.

Method	PSNR↑	PSNR-Y↑	SSIM↑	ΔE_{ITP} ↓	HDR-LPIPS↓	HDR-VDP-3↑	FR-HIDRO↓
BT.2446c	<u>33.16</u>	<u>34.26</u>	<u>0.8442</u>	**37.83**	0.148	**7.397**	0.660
Reinhard operator	31.48	32.24	0.8270	43.59	0.155	<u>7.173</u>	0.662
HDRTVNet++	31.53	32.05	0.8125	44.92	0.147	7.144	0.671
FMNet	31.64	32.11	0.8137	44.85	0.147	7.135	0.675
KUNet	31.51	31.96	0.8078	45.26	0.143	7.058	0.647
ITM-LUT	31.46	32.06	0.8085	45.49	<u>0.142</u>	7.090	<u>0.639</u>
LEDiff	18.30	18.99	0.6480	103.15	0.496	4.723	0.898
X2HDR	23.78	24.69	0.7578	51.28	0.366	5.532	0.758
PromptIR	29.21	29.67	0.6756	53.11	0.185	6.141	0.690
VAE + RQS (control)	26.05	27.33	0.7231	91.70	0.237	5.814	0.863
LumaFlux	**33.31**	**35.17**	**0.8480**	<u>40.19</u>	**0.138**	6.919	**0.544**

Table 11: The same seven metrics on the native-pair track of 117 HDRTV1K test pairs, scored with the mixed-corpus LumaFlux checkpoint. Every displayed method has a complete 117-row score file with exact ID accounting; ITM-LUT has no result on this track. Table 7 scores the same pairs with its five-metric legacy suite and the HDRTV1K-specific checkpoint.

Method	PSNR↑	PSNR-Y↑	SSIM↑	ΔE_{ITP} ↓	HDR-LPIPS↓	HDR-VDP-3↑	FR-HIDRO↓
BT.2446c	25.00	24.89	0.9511	39.64	0.078	7.632	0.514
Reinhard operator	24.49	24.32	0.9342	43.97	0.083	7.637	0.485
HDRTVNet++	**38.11**	**40.06**	**0.9824**	8.13	<u>0.043</u>	<u>8.224</u>	<u>0.338</u>
FMNet	<u>38.00</u>	<u>40.05</u>	<u>0.9822</u>	<u>8.10</u>	0.047	8.188	0.358
KUNet	37.99	39.85	0.9820	**7.84**	**0.038**	**8.235**	**0.258**
LEDiff	10.99	11.28	0.5245	198.58	0.593	3.964	1.088
X2HDR	18.65	18.80	0.6584	92.20	0.445	5.071	0.874
PromptIR	23.18	23.02	0.8617	53.35	0.099	7.554	0.475
VAE + RQS (control)	21.47	23.30	0.9215	71.86	0.201	6.192	0.804
LumaFlux	23.56	27.71	0.9316	47.59	0.170	6.899	0.632

Tables 10 and 11 keep the two real-pair tracks apart instead of pooling unlike SDR formation processes; every listed method satisfies exact prediction and score accounting on each track it enters (ITM-LUT has no native-track result). Table 11 and the in-domain LumaFlux row of Table 7 evaluate the same 117 published HDRTV1K test pairs, but they are not interchangeable: the legacy comparison uses its five-metric suite and the HDRTV1K-specific checkpoint, whereas the native-pair track uses the seven-metric Luma-Eval suite and the mixed-corpus checkpoint.

The pooled CRF diagnostic in Table 12 cannot identify a causal compression slope because its two column groups contain different TMO mixtures. At CRF 39 LumaFlux has the highest PSNR and



Figure 17: Headroom maps of five lifts, each at its stored keyframe with the largest area above the 100-nit SDR ceiling: gray is luminance up to the ceiling, color is above it. Of the pixels above it, 90.4% to 98.3% lie among the equally many brightest source pixels.

the lowest color error of the ranked methods. The VAE + RQS control stays at or below 20.53 dB, so its flatter gap is not a robustness win at that absolute quality.

In the fixed-QP study, the HDRTV1K frame set and codec path stay constant, so QP 42 minus QP 27 measures degradation on matched inputs. Figure 9 in the main text plots this change; Table 13 lists all values. Absolute quality and compression sensitivity must be read separately: HDRTVNet++ starts at 35.90 dB and remains highest at 33.13 dB, despite losing 2.77 dB. The mixed-corpus LUMAFLUX row is the flattest, losing 0.55 dB, but lies below both CNN rows. The HDRTV1K-specific checkpoint raises LUMAFLUX to 33.94 and 32.33 dB at the two end points, a 1.61 dB loss. These are means over the same images, without repeated-training uncertainty.

Table 12: Pooled CRF diagnostic on the synthetic track. The two CRF groups contain different registered TMO mixtures, so each gap is descriptive and not a compression slope. Bold and underline rank the actual methods; the gray VAE + RQS control is not ranked.

	PSNR↑ (dB) at CRF			ΔE_{ITP} ↓ at CRF		
Method	23	39	gap↓	23	39	gap↓
BT.2446c	**32.38**	18.63	13.76	**28.66**	81.09	52.43
Reinhard	31.85	18.75	13.10	29.92	80.61	50.69
HDRTVNet++	27.88	20.15	7.73	48.65	77.34	28.69
FMNet	27.46	20.14	**7.32**	51.52	77.63	**26.12**
ITM-LUT	27.68	20.10	7.58	49.63	77.93	28.31
VAE + RQS	20.53	17.36	3.17	100.49	118.29	17.81
LUMAFLUX	30.71	**20.39**	10.32	39.34	**66.72**	27.37

Table 13: Four metrics on the same 117 HDRTV1K pairs after intra-only HEVC coding of the SDR input at four QPs; Δ is QP 42 minus QP 27, so a Δ closer to zero means less degradation. Bold and underline mark the best and second-best method in each metric block and column. The LUMAFLUX row is the mixed-corpus checkpoint and the (in-domain) row the HDRTV1K-specific one.

Metric	**Method**	QP 27	QP 32	QP 37	QP 42	Δ
PSNR↑ (dB)	HDRTVNet++	**35.90**	**35.30**	**34.34**	**33.13**	−2.77
	FMNet	35.56	34.97	34.04	32.85	−2.71
	LUMAFLUX	27.29	27.27	27.04	26.75	**−0.55**
	LUMAFLUX (in-domain)	33.94	33.67	33.12	32.33	−1.61
SSIM↑	HDRTVNet++	0.9645	0.9561	0.9403	**0.9150**	−0.0495
	FMNet	**0.9647**	**0.9563**	**0.9404**	0.9149	−0.0498
	LUMAFLUX	0.9309	0.9261	0.9142	0.8925	**−0.0384**
	LUMAFLUX (in-domain)	0.9430	0.9370	0.9243	0.9021	−0.0409
ΔE_{ITP} ↓	HDRTVNet++	**10.41**	**11.18**	**12.47**	**13.98**	3.58
	FMNet	10.53	11.33	12.62	14.18	3.65
	LUMAFLUX	30.32	30.56	31.30	32.08	**1.76**
	LUMAFLUX (in-domain)	13.64	14.09	15.01	16.20	2.56
HDR-VDP-3↑	HDRTVNet++	**8.409**	**8.167**	**7.759**	**7.198**	−1.211
	FMNet	8.329	8.097	7.704	7.156	−1.173
	LUMAFLUX	7.615	7.525	7.308	6.948	**−0.667**
	LUMAFLUX (in-domain)	7.918	7.791	7.533	7.117	−0.802

The protocol follows the DIDNet robustness study (Xu et al., 2023), adapted to the staged still-image release. DIDNet uses 87 seven-frame SDR windows per QP, 348 pairs in total. The staged HDRTV1K test release exposes 117 still PNG pairs, so each SDR image is encoded independently as one intra-only HEVC frame, giving 117 pairs per QP and 468 in total. SR-SIM remains available in the machine-readable summary files.

G LUMA-EVAL BY TMO FAMILY

Table 14 splits the aggregate of Table 1 by degradation family. A baseline appears only when its prediction and score accounting is exact and complete; partial or superseded runs are excluded.

Each baseline class leads where its assumption holds: the analytic inverses on the standards they invert, and the CNNs on global curves. LUMAFLUX is second on the standards and has the lowest production color error; it is the only method ranked in the top two in more than one family, and it has the best pooled PSNR (24.23 dB) and color error (56.54).

Table 14: Luma-Eval by TMO family on the 100 fixed synthetic-track IDs. Each column pair reports PU21-PSNR and ΔE_{ITP}, macro-averaged from aligned per-frame scores over the video-variant groups of that family; the last pair repeats Table 1. Bold and underline mark the best and second-best value per column.

	Standards		Global curves		Production		All	
Method	PSNR↑	ΔE_{ITP} ↓	PSNR↑	ΔE_{ITP} ↓	PSNR↑	ΔE_{ITP} ↓	PSNR↑	ΔE_{ITP} ↓
BT.2446c	**30.95**	**36.55**	20.54	63.34	18.14	89.03	<u>23.74</u>	<u>61.60</u>
Reinhard operator	29.53	39.90	21.52	64.37	**18.57**	<u>84.78</u>	23.62	61.76
HDRTVNet++	28.32	45.77	**24.77**	**37.92**	15.44	115.11	23.03	66.67
FMNet	28.03	47.84	<u>24.71</u>	<u>38.81</u>	15.34	115.73	22.86	67.92
ITM-LUT	28.18	46.21	24.67	39.14	15.35	115.78	22.92	67.41
VAE + RQS (control)	21.39	92.69	22.99	58.87	11.49	178.26	18.54	111.67
LUMAFLUX	<u>30.16</u>	<u>39.78</u>	22.91	49.59	<u>18.50</u>	**81.76**	**24.23**	**56.54**

H LUMA-EVAL FRAME WIN RATES

Table 15 counts the percentage of aligned frames on which each method is best. Exact ties split one frame equally, and only methods covering all 100 fixed IDs enter. LUMAFLUX wins most frames on PSNR and on all three perceptual metrics, by a wide margin on the latter (62.0 to 69.2%); the Reinhard operator wins most frames on SSIM and color error.

Table 15: Luma-Eval frame win rates (%): the share of aligned frames on which each method is best, with exact ties split equally. Bold and underline mark the best and second-best rate per metric.

Method	PSNR	SSIM	ΔE_{ITP}	LPIPS	VDP-3	HIDRO
BT.2446c	14.2	<u>19.1</u>	16.5	7.2	<u>13.2</u>	6.5
Reinhard operator	<u>27.2</u>	**37.1**	**27.5**	<u>14.2</u>	11.2	3.5
HDRTVNet++	6.2	11.1	11.0	1.2	6.0	<u>17.0</u>
FMNet	13.2	13.1	14.0	0.2	1.2	3.0
ITM-LUT	3.2	0.1	3.0	7.2	2.2	1.0
VAE + RQS (control)	7.0	4.1	2.0	1.0	4.0	1.0
LUMAFLUX	**29.2**	15.1	<u>26.0</u>	**69.2**	**62.0**	**68.0**

I PER-VARIANT LUMA-EVAL RESULTS

Table 16 gives the complete LUMAFLUX decomposition behind the macro-average in Table 1. The unequal group counts arise from manifest availability and are preserved rather than imputed. The split follows compression first: every CRF-23 variant scores at least 24.75 dB, and every CRF-39 variant at most 23.04 dB.

Table 16: Per-variant LUMAFLUX results on the synthetic track, grouped by the CRF of the registered variant. Each row is a macro-average over N video groups. Bold and underline mark the best and second-best value per metric.

Variant	N	PSNR↑	PSNR-Y↑	SSIM↑	ΔE_{ITP} ↓	HDR-LPIPS↓	HDR-VDP-3↑	FR-HIDRO↓
CRF 23								
BT.2390 EETF + gamut map	20	**32.76**	**32.94**	<u>0.9218</u>	39.60	**0.102**	**7.140**	0.426
BT.2446 Method C + gamut map	4	<u>27.83</u>	<u>28.49</u>	0.9007	**25.56**	0.134	<u>6.783</u>	<u>0.401</u>
Hard clip + gamut map	1	25.12	25.27	0.9134	<u>39.48</u>	<u>0.117</u>	6.769	**0.305**
Reinhard	4	24.75	26.05	**0.9333**	51.81	0.173	5.908	0.578
CRF 39								
BT.2446 Method A	6	23.04	24.03	0.8760	49.84	0.282	5.577	0.638
Gamma clip	17	22.34	22.61	0.8300	49.66	0.306	5.508	0.673
OpenColorIO v2	6	18.87	18.68	0.7864	82.87	0.349	4.931	0.749
YouTube LogC	20	18.39	18.69	0.6960	81.43	0.431	4.817	0.835

A cross-operator stress test isolates degradation specificity over 20 frames per direction. BT.2446c applied to Reinhard SDR reaches 27.21 dB and 26.93 ΔE_{ITP}; the Reinhard operator applied to BT.2446c SDR reaches 28.73 dB and 24.32. LUMAFLUX reaches 31.34/19.58 and 31.04/18.84 on the corresponding Reinhard and BT.2446c sets. This collapse under mismatched operators motivates

learning across the full degradation family. Because the sample is small, we use it as mechanism evidence, not a benchmark claim.

J EFFICIENCY AND MODEL SCALE

Table 17 reports measured export latency and model-scale evidence under each method's registered protocol. Latency comes from complete exports at 1080p, except LEDiff and X2HDR, which use their registered 512^2 protocol. Parameter counts come from released weights, TFLOPs from operator-level traces of the real models at their registered resolutions, and peak memory from the common measurement artifact. The KUNet latency is the expert-track export timer; the synthetic run did not persist one. LUMAFLUX trains 71.32 M of its 12.5 B parameters, fewer than either generative baseline, and its eight-step export costs about one fifth of the 40-step schedule.

Table 17: Efficiency comparison under each method's registered export protocol. Latency comes from complete 1080p exports, except LEDiff and X2HDR at their registered 512^2 protocol; parameters come from released weights, TFLOPs from operator-level traces, and peak memory from the common measurement artifact. The two LUMAFLUX rows are the 8-step and 40-step samplers. BT.2446c is analytic; *ITM-LUT runs in a custom compiled kernel that the tracer cannot observe.

Method	Steps	Params (M)	Train. (M)	s/frame↓	TFLOPs↓	Mem. (GB)↓
BT.2446c	1	0.00	0.00	0.415	–	–
Reinhard operator	1	0.00	0.00	0.393	–	–
HDRTVNet++	1	0.59	0.59	0.857	0.5	2.31
FMNet	1	1.30	1.30	0.640	1.5	2.16
KUNet	1	2.16	2.16	0.557	3.7	3.76
ITM-LUT*	1	0.60	0.60	0.367	–	0.18
PromptIR + Reinhard	1	35.59	35.59	1.708	10.9	17.86
VAE + RQS (control)	1	–	–	1.175	–	–
LEDiff	50	1370.23	859.53	4.449	149	6.37
X2HDR	30	17020.02	149.42	4.244	1,023	33.77
LUMAFLUX	8	12484.77	71.32	5.282	2,067	27.06
LUMAFLUX	40	12484.77	71.32	24.993	10,100	27.06

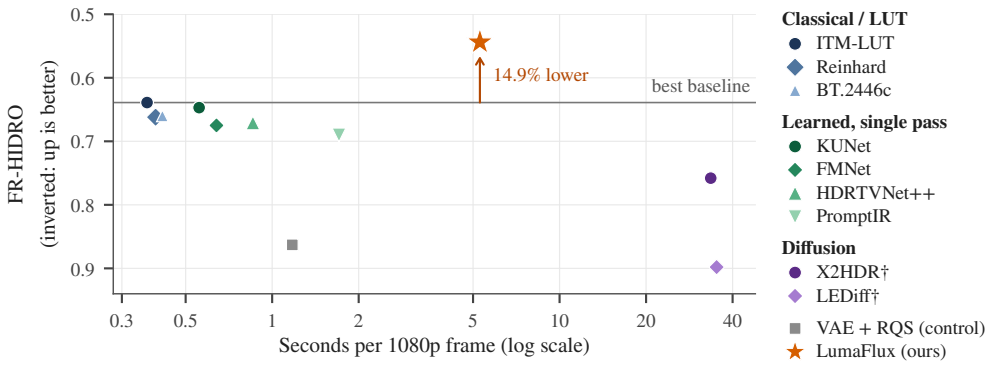

Figure 18: FR-HIDRO against measured latency, expert track (Table 17: one GH200, batch size 1; axis inverted, up is better). †Timed at 512^2, plotted at the pixel-scaled 1080p time.

Figure 18 places the same latencies against Luma-Eval FR-HIDRO on the expert track, where all 11 timed methods are scored (Table 10).

The single-pass baselines take 0.367 to 1.708 s per frame and score 0.639 to 0.690. Eight-step LUMAFLUX reaches 0.544, 14.9% below the best baseline (ITM-LUT, 0.639), at 5.282 s per frame, 6.2 times the cost of HDRTVNet++, so every single-pass method is faster. On the synthetic track (Table 1) the single-pass baselines take 0.367 to 0.857 s and score 0.703 to 0.761, and LUMAFLUX reaches 0.631, 10.2% below the best of them. LEDiff and X2HDR were timed at 512^2 (4.449 and 4.244 s); scaled by pixel count to 1080p these become 35.19 and 33.57 s, 6.7 and 6.4 times the LUMAFLUX time, and they score 0.898 and 0.758 on the expert track. The 40-step schedule costs

4.7 times the eight-step one (24.993 s, Table 17) and did not improve the six-frame calibration, so every reported LUMAFLUX result uses eight steps.

J.1 GENERATIVE MODEL OPERATING POINTS

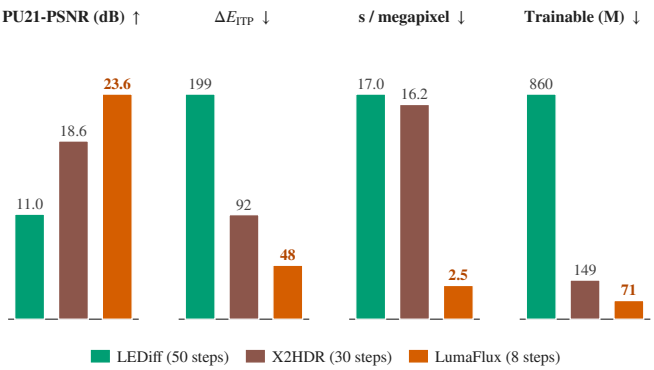

Figure 19: For each generative model, quality is measured on the complete native track; cost is export time divided by output pixels at each registered protocol. LEDiff and X2HDR export at 512^2, LUMAFLUX at 1080p, so the per-pixel cost is descriptive and not a matched-resolution speed. Trainable parameters and sampler steps repeat Table 17.

The common native track gives LUMAFLUX 23.56 dB PU21-PSNR against 18.65 for X2HDR and 10.99 for LEDiff. The measured exports use different resolutions, so Figure 19 should be read together with the raw timing and resolution records of Table 17.

K ADDITIONAL QUALITATIVE DIAGNOSTICS

Figures 21, 22, and 23 give complementary spatial views of the comparison scenes: diagonal preview-and-error maps, enlarged registered regions, and log-luminance patches for every completed method. Figure 24 follows luminance along one image row: the SDR input stays at or below its 100-nit peak, and LUMAFLUX follows the reference profile most closely in both scenes, with a mean absolute log deviation of 0.058 and 0.078 against 0.166 to 0.272 for BT.2446c and HDRTVNet++. The compact comparison in the main paper is Figure 10: for each scene it shows the full SDR input with a patch box and that patch for each method, with the full-frame PU21-PSNR of the frame above each patch. Figure 20 re-renders one stored result at several display exposures.

K.1 MATCHED DISPLAY-EXPOSURE EXAMPLES

Figure 20 uses one selected city-sunset scene with a clean-compression CRF-23 input. Every column multiplies decoded linear radiance by 2^{EV} before the shared filmic display transform. The crop, color conversion, and mapping are identical for every method. HDRTVNet++ is the strongest CNN by mean PU21-PSNR on the complete synthetic grid. Here LUMAFLUX retains a warmer sky and shadowed-building rendition closer to the reference. Selection favors an illustrative scene; aggregate conclusions use the complete evaluation tables.

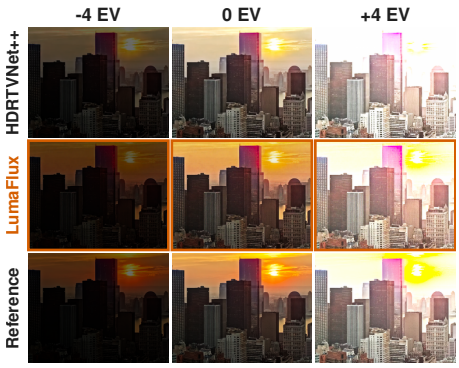

Figure 20: One stored crop at matched display exposures of -4, 0, and $+4$ EV; the display mapping is shared.

Negative display exposure reveals bright-region structure, while positive exposure makes shadow differences visible. These views probe the appearance of a fixed reconstructed signal across display settings; they do not measure robustness to altered model inputs. No model is rerun, and no method receives a separate display normalization. The source archive records the frame identifier, exact integer crop, source hashes, and shared transform; it also retains the full five-stop source asset.

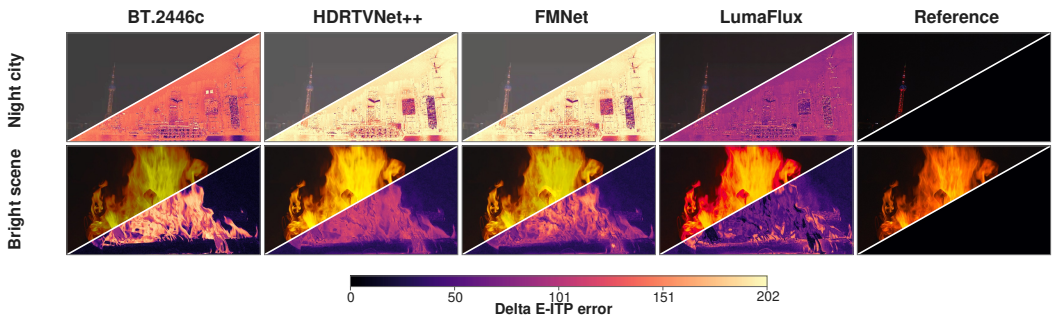

Figure 21: Each panel splits diagonally into the rendered preview and its per-pixel ΔE_{ITP} error on a shared scale. The maps localize color error around the waterfront and flame cores. LUMAFLUX reduces error in many selected regions, while its flame rendition retains visible chroma overshoot.

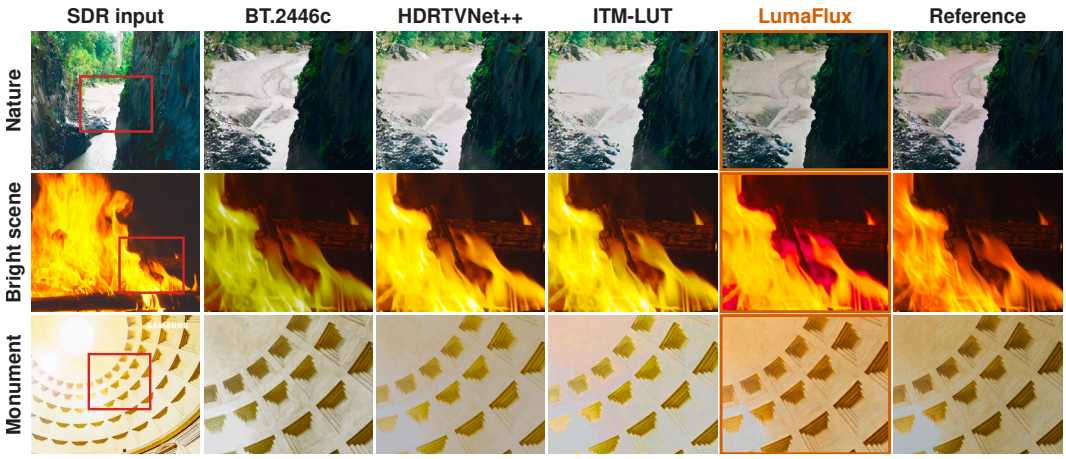

Figure 22: The red box on each SDR frame marks the region shown at larger scale for every HDR method and the reference, all under one fixed display transform; LUMAFLUX is outlined in its plot color. These are the three comparison scenes of the main paper. The flame row also exposes residual chroma overshoot in LUMAFLUX.

L EVALUATION PROTOCOL DETAILS

HDRTV4K is not included because its ground-truth test package could not be obtained through an open channel. NTIRE and AIM inverse-tone-mapping sets are outside this benchmark because they target relative, scene-referred linear reconstruction rather than the absolute display-referred PQ BT.2020 signal evaluated here (Wang et al., 2025a).

Each cited HDRTV1K baseline cell in Table 7 is traceable to a published table or an official repository release. Methods without first-party HDRTV1K results (Deep SR-ITM, FMNet) use the retrained values reported in the HDRTVNet and HDRTVNet++ comparisons (Chen et al., 2021b; 2023); the ITM-LUT row uses its official repository's HDRTV1K checkpoint; and the RealRep row follows that paper's unknown-degradation protocol, so it measures generalization rather than the standard split. We do not substitute a run from a different implementation when a published value exists. Unreported metrics are marked "–." ICTCPNet is absent because a published value under this protocol could not be verified.

In Luma-Eval, KUNet and PromptIR intersect 86 of the 100 primary IDs (73 of 78 groups), so they are omitted from the primary synthetic table and retained only in complete track-specific comparisons. The run-by-us synthetic comparison fixes 100 frame IDs, color conversion, frame selection, and metric implementation. Completed full-manifest results are projected onto those IDs. LEDiff and X2HDR use the disclosed five-variant subset and the common evaluation luminance anchor described in Section 6; they enter common-grid analyses only when every required ID is present. PromptIR remains a general-purpose restoration comparison. LUMAFLUX uses the same inputs and the same output metadata. The expert and native tables retain their separate 100-pair and 117-pair

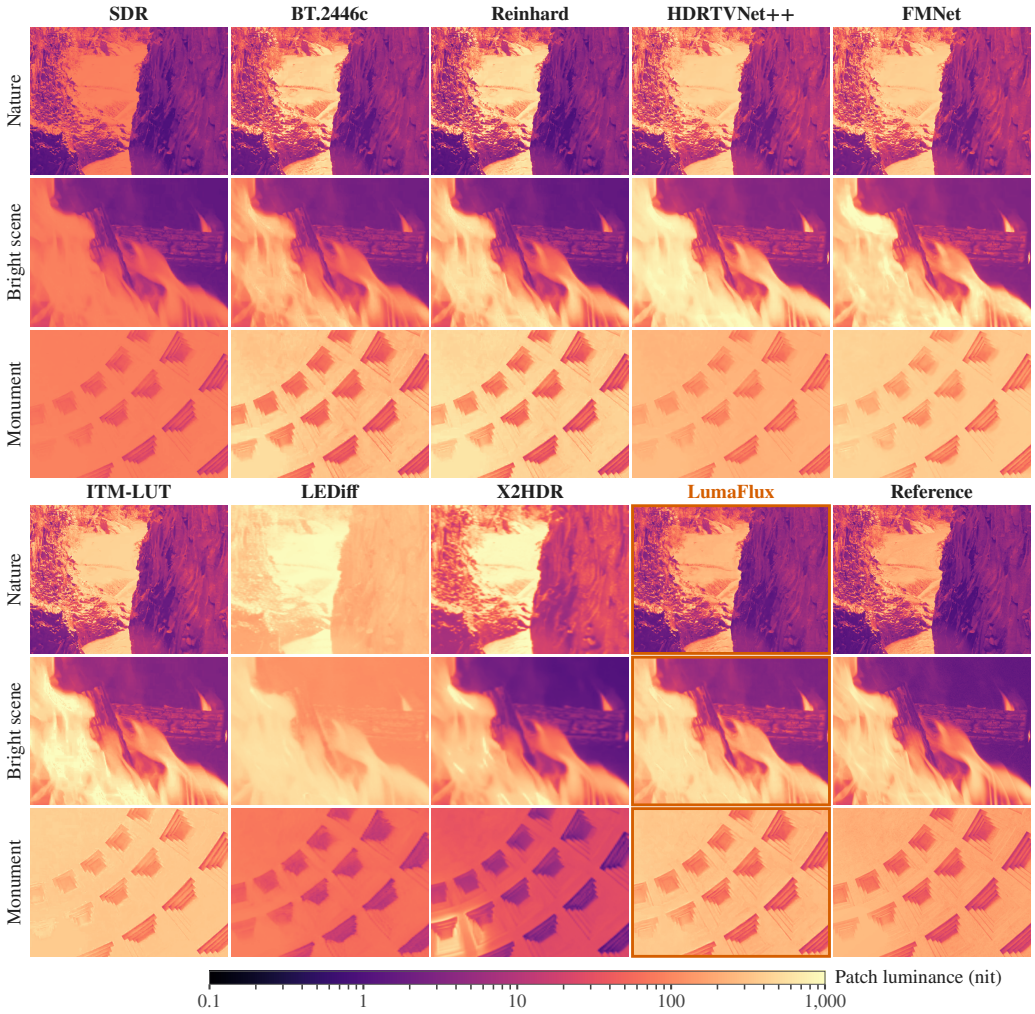

Figure 23: The upper and lower blocks repeat the same three registered regions as log-luminance patches for every method, center-cropped to 4:3, with five columns per block. All panels share one logarithmic scale from 0.1 to 1,000 nit; SDR uses a 100-nit display white, and LUMAFLUX is outlined. These are selected qualitative examples, not an aggregate ranking.

manifests. HDR-LPIPS applies the AlexNet LPIPS network to PU21 values scaled to $[-1, 1]$, and every reported difference between methods is computed from unrounded values.

For the temporal metric, a predicted sequence and its HDR reference are both mapped to PU21 and aligned using the frozen evaluation pipeline. Signed excess flicker energy is the temporal residual energy of the prediction minus the reference temporal energy under the same alignment and masking rules. Table 8 takes the absolute value per clip and then the unweighted five-clip mean, so excess variation and over-smoothing are both penalized. The reference row is zero by definition. Figure 15 uses only traces marked complete in the landed manifest; unavailable LUMAFLUX traces are skipped rather than reconstructed from single midpoint frames.

M INFERENCE PROCEDURE

Algorithm 1 summarizes prompt-free inference with the frozen Flux MM-DiT $\mathcal{F}_{\boldsymbol{\theta}}$, PGA and PCM adapters, HDR Residual Coupler h_ψ, RQS head f_{RQS}, and conditioner Ψ. The solver states map to the notation of Section 4 as $\mathbf{z}^{(0)} = \mathbf{z}_1$ and $\mathbf{z}^{(K)} = \mathbf{z}_0$. Every reported result uses $K = 8$ steps.

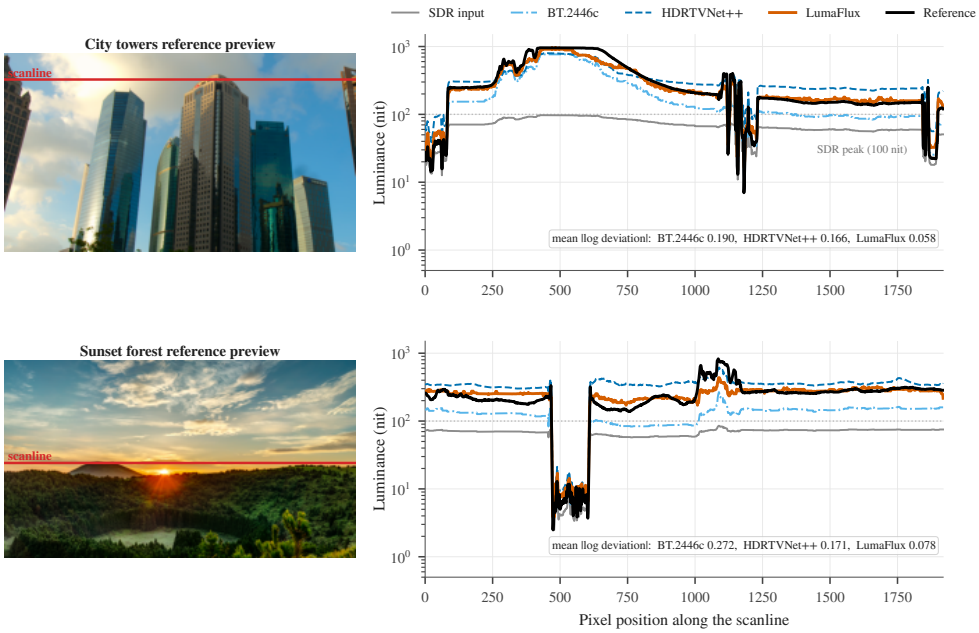

Figure 24: For two scenes, the left panel shows the reference preview with the sampled row in red, and the right panel plots luminance along that row on a logarithmic nit axis for the SDR input, BT.2446c, HDRTVNet++, LUMAFLUX (mixed-corpus checkpoint), and the reference. The mean absolute log-luminance deviation from the reference is 0.190, 0.166, and 0.058 on City towers and 0.272, 0.171, and 0.078 on Sunset forest for BT.2446c, HDRTVNet++, and LUMAFLUX. These are selected rows, not an aggregate ranking.

Algorithm 1 Prompt-free SDR-to-HDR conversion

Require: SDR frame x_{sdr}; frozen $\mathcal{F}_{\boldsymbol{\theta}}$, VAE, SigLIP; adapters, h_ψ, f_{RQS}, Ψ; $K = 8$ steps; shared noise $\boldsymbol{\epsilon}$

Convert x_{sdr} to linear BT.2020; compute luminance, gradient, saturation, global, and spectral statistics and frozen SigLIP features to form conditioning $\mathbf{c}$
$\mathbf{z}_{\mathrm{sdr}} \leftarrow \mathrm{VAE}_{\mathrm{enc}}(x_{\mathrm{sdr}})$; $\mathbf{z}^{(0)} \leftarrow \mathbf{z}_{\mathrm{sdr}} + 0.05\,\boldsymbol{\epsilon}$ ▷ Reuse $\boldsymbol{\epsilon}$ across video frames
for $k = 0, \ldots, K-1$, with $t_k = 1 - k/K$ **do**
    for each MM-DiT block ℓ **do**
        Obtain timestep-layer gates from $\Psi(t_k, \ell)$
        Apply PCM before the frozen block using Equations 7 and 8
        Evaluate the frozen block with PGA in attention (Equation 6)
        Add the HDR residual coupler after the block (Equation 9)
    end for
    $\mathbf{z}^{(k+1)} \leftarrow \mathbf{z}^{(k)} + (t_{k+1} - t_k)\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{z}^{(k)}, t_k, \mathbf{c})$ ▷ Adapted velocity; descending time
end for
$x_{\mathrm{out}} \leftarrow \mathrm{VAE}_{\mathrm{dec}}(\mathbf{z}^{(K)})$; interpret as PQ BT.2020 RGB and convert to Y′CbCr luma and chroma
Predict RQS parameters; apply video EMA when a previous-frame state exists
Apply monotone RQS to luma; refine chroma; convert back to PQ BT.2020 RGB
return $\widehat{x}_{\mathrm{hdr}}$

N LIMITATIONS AND BROADER IMPACT

The frozen VAE was not trained as an HDR codec, so the RQS head calibrates its output but cannot recover information that is absent from heavily clipped or band-limited SDR. Temporal stabilization is inference-time only and does not model long-range motion. Across the five selected qualitative scenes, LUMAFLUX has mean ITP chroma 60.65 versus 49.80 for the matched references, exposing measurable oversaturation rather than hiding it through method-specific display tuning. The primary synthetic comparison contains 20 contents and 100 frame pairs. Expert and native results are reported separately under exact completion checks. HDRTV4K is absent because its test package

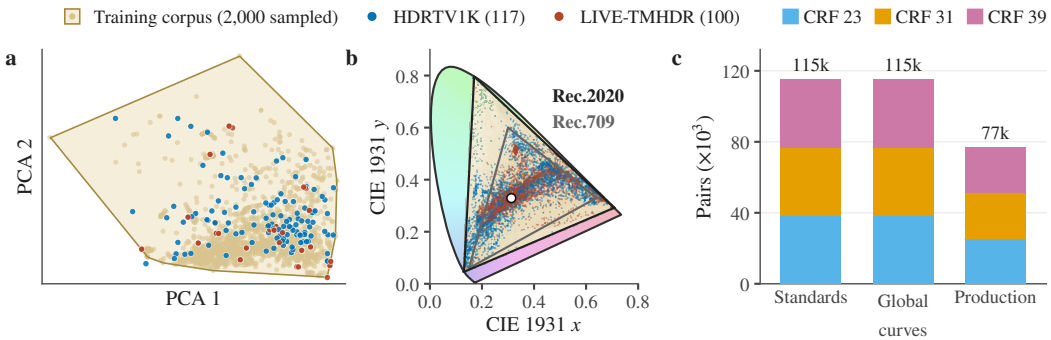

Figure 25: (a) PCA of ten per-frame statistics for 2,000 frames sampled from the 314,396 training pairs (sand points with their convex hull), the 117 HDRTV1K test frames (blue), and 100 LIVE-TMHDR frames (rust); the corpus hull contains every LIVE-TMHDR frame and 116 of the 117 HDRTV1K frames. (b) CIE 1931 xy chromaticity: the sand footprint bins every stored corpus sample and fills the BT.2020 triangle, with benchmark samples overlaid in the same colors. (c) Training pairs by tone-mapping family and CRF (307,752 registered synthetic rows; the full paired corpus has 314,396 rows).

is Baidu-gated. The aggregate ranking therefore needs confirmation on broader tracks. The six-frame solver study is a calibration, not a baseline runtime comparison. Deployment requires correct display metadata and EOTF settings. A controlled subjective study is planned.

LumaFlux may help remaster legacy SDR archives and user-generated content for HDR presentation. Creative SDR looks can be intentional, so deployment should preserve an unmodified source and offer a bypass. The dataset release separates redistributable open content from scripts and IDs for restricted sources.

O Corpus Construction and Signal Normalization

Training and evaluating inverse tone mapping requires physically valid HDR targets and SDR inputs that reflect production pipelines. The training corpus contains 314,396 SDR-HDR pairs curated from 2,092 source videos. Every record identifies its source frame, color metadata, degradation parameters, integer storage type, and checksum, so the training set is an auditable product of the curation code. Table 6 reports decoded-frame statistics for every source, and Figure 25 compares corpus diversity against the evaluation benchmarks.

O.1 Training Corpus

The user-generated (UGC) source pool contains 856 CHUG reference contents (Saini et al., 2025), of which 846 enter training and ten are reserved. These videos contribute handheld motion, exposure variation, mobile capture characteristics, and platform compression. Content IDs are retained so the split can be reconstructed without redistributing restricted media.

The professional (PGC) source pool contains all 40 LIVE-TMHDR references and 120 associated expert-graded SDR clips (Venkataramanan & Bovik, 2024), 1,235 native HDRTV1K training pairs (Chen et al., 2021b), and the HDR10 master of Netflix Open Content's *Sol Levante* under CC BY (Netflix, 2026). LIVE-TMHDR references contain one extra leading frame, so expert frame n is paired with reference frame $n + 1$. This offset is encoded in the manifest.

The curated manifest contains 171,384 CHUG examples and 143,012 PGC examples; the shard-backed training loader shuffles these rows. The corpus includes native HDRTV1K training pairs. We distinguish the 100,000-step mixed-corpus checkpoint from the separately trained 50,000-step HDRTV1K-specific checkpoint. When a source lacks an SDR counterpart, the SDR input is generated through

$$x_{\mathrm{sdr}} = \mathcal{Q}_{\mathrm{codec}} \circ \mathcal{M}_{2020\to709} \circ \mathcal{T}_{\mathrm{tone}}(x_{\mathrm{pq}}; \theta_{\mathrm{tone}}), \tag{15}$$

where $\mathcal{T}_{\mathrm{tone}}$ samples eight tone-mapping operators, $\mathcal{M}_{2020\to709}$ converts the gamut, and $\mathcal{Q}_{\mathrm{codec}}$ applies three CRF settings. The frozen curation recipe fixes each source's decode matrix before any

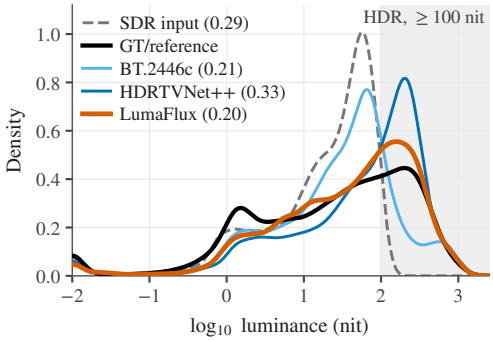

Figure 26: Pooled luminance distributions on the 100-frame evaluation set. Legend values are Wasserstein distances to the reference in $\log_{10}$ nit (lower is better).

tone or gamut operation, and the manifest, source specification, and code regenerate the full corpus in approximately four CPU hours.

O.2 Luma-Eval and Release Policy

Luma-Eval has three tracks: synthetic tone-mapping and codec degradations, expert-graded SDR from LIVE-TMHDR, and native paired SDR/HDR. The primary synthetic comparison fixes 100 aligned frame IDs from 20 held-out CHUG and LIVE-TMHDR contents, and its `ALL` row is a macro-average over 78 video-variant groups. Methods scored on a larger 2,500-pair manifest are projected onto these exact 100 IDs before comparison. LEDiff and X2HDR cover five disclosed synthetic variants and enter common-grid analyses only when every required ID is present. The expert and native tracks contain 100 and 117 pairs, respectively, and a method enters a track table only when its predictions and scores are complete for every pair. Figure 26 pools per-pixel luminance over the 100 evaluation frames: the analytic operator largely retains the SDR pile-up below the 100-nit ceiling, the CNN concentrates its mass in a narrow band just above it, and LumaFlux matches the reference's broad highlight distribution most closely.

HDR sources are decoded according to their mastering metadata, transformed into BT.2020 primaries, and stored in PQ. If p_i and Γ_i denote the native primaries and transfer function of source i, normalization computes

$$x_{\mathrm{pq}}^{(i)} = \Gamma_{\mathrm{OETF}}^{\mathrm{PQ}}\left(\mathcal{M}_{p_i \to 2020}\left((\Gamma_i)^{-1}(x_{\mathrm{hdr}}^{(i)})\right)\right). \tag{16}$$

The release records mastering metadata and conversion parameters per content. Restricted media are represented by deterministic curation scripts, content IDs, split manifests, and checksums.

P Supplementary Related Work

Analytic ITM expands SDR luminance with a prescribed curve and saturation rule. Reinhard et al. (2002) derive a photographic tone-reproduction operator, while ITU BT.2446 (ITU-R, 2021) specifies broadcast conversion procedures. These methods are reproducible and inexpensive, but a fixed curve cannot recover content-dependent structure removed by tone mapping, gamut compression, and quantization.

For SDRTV-to-HDRTV conversion, HDRTVNet++ (Chen et al., 2023), Deep SR-ITM (Kim et al., 2019), HDCFM (He et al., 2022), FMNet (Xu et al., 2022), KUNet (Wang et al., 2022), ICTCP-Net (Huang et al., 2023), and HDRTVDM (Guo et al., 2023a) learn paired display-referred mappings. Beyond-FMG (He et al., 2025) adds real-HDRTV priors, while RealRep (Xu et al., 2026) separates luminance and chrominance across SDR styles. We retain author-reported results under their stated protocols; HyCondITM (Shao et al., 2022) remains citation-only because it lacks released weights.

Video-oriented ITM adds trained spatiotemporal modules or learned filtering stages (Xu et al., 2023; Kim et al., 2020), and post-hoc temporal filtering adds a stage after conversion; framewise CNN

converters (Chen et al., 2023; He et al., 2022; Xu et al., 2022) have no temporal state and can carry input fluctuation into independently estimated HDR mappings. LUMAFLUX stabilizes the trained image model without temporal architecture or retraining, coupling deterministic per-frame transport through shared bridge noise and smoothing only the applied tone parameters at inference.

HDRCNN (Eilertsen et al., 2017), SingleHDR (Liu et al., 2020), HDRUNet (Chen et al., 2021a), and GlowGAN (Wang et al., 2023) reconstruct relative or scene-referred HDR radiance from camera images. Their output domain differs from the absolute, display-referred PQ BT.2020 signal targeted here, so they frame single-image HDR recovery but do not enter our television benchmark.

Among generative adaptations in perceptual space, X2HDR (Wu et al., 2026) is the closest methodological relative. It freezes the FLUX.1-dev VAE and adapts the denoiser with LoRA in PU21 space, supporting text-to-HDR generation and RAW-to-HDR reconstruction after per-image rescaling to a 4,000-nit peak. This result motivates our use of a frozen generative prior and makes X2HDR a strong run-by-us baseline. Our contribution beyond perceptual encoding is the coordinated physical and semantic conditioning of a frozen backbone, a monotone tone-field decoder, and inference-time video stabilization. LUMAFLUX targets paired SDR-video restoration under a fixed 1,000-nit PQ BT.2020 convention and explicitly evaluates temporal fidelity; the common evaluation wrapper used for X2HDR is specified in Section 6.

LEDiff (Wang et al., 2025b) fuses generated exposure latents to recover linear HDR, and LumaGuide (Chen et al., 2026) shapes the sampling distribution of a pretrained diffusion model to generate HDR without retraining. PromptIR (Potlapalli et al., 2023) conditions a general image-restoration network rather than targeting ITM. FlashVSR (Zhuang et al., 2025) targets streaming video super-resolution rather than inverse tone mapping and is therefore omitted from quantitative comparisons. LoRA (Hu et al., 2022) and adapter methods (Chen et al., 2022; Mou et al., 2024) provide the broader basis for parameter-efficient backbone specialization. RB-Modulation (Rout et al., 2025) controls style and content through a reference-feature terminal cost and attention aggregation at inference. Our physical and semantic adapters are trained for paired HDR fidelity; their modulation does not optimize a reference-style objective.

The AIM 2025 ITM challenge (Wang et al., 2025a) studies single-image recovery from synthetic exposure and camera-response degradations into scene-referred linear HDR, with evaluation in PU21 space. Our target is a display-referred PQ BT.2020 master with fixed absolute luminance, real expert grades, codec variants, and video consistency. The challenge winners provide evidence about modern ITM architectures, but their reported scores are not directly protocol-compatible. We retain these results under their stated protocols; evaluating their predictions here would require an explicit conversion into the common output convention.

In concurrent HDR-video work, DiffHDR (Yu et al., 2026) casts LDR-to-HDR video as generative radiance inpainting in log-gamma space with a pretrained video diffusion model. HDR Video Generation via Latent Alignment (Korem et al., 2026) uses logarithmic encoding to align HDR data with a pretrained video model for lightweight generative adaptation. Single-Shot HDR Recovery (Talegaonkar et al., 2026) treats one-image reconstruction as conditional video generation of an exposure bracket followed by learned fusion. Generating HDR Video from SDR Video (Tedla et al., 2026) predicts exposure-bracketed linear video and merges it into HDR. These concurrent systems strengthen the case for generative priors, while our focus remains faithful display-referred restoration under a common paired protocol.

HDRTV1K (Chen et al., 2021b) and HDRTV4K (Guo et al., 2023a) provide established paired benchmarks. CHUG (Saini et al., 2025) contributes varied user-generated HDR, and LIVE-TMHDR (Venkataramanan & Bovik, 2024) includes professionally produced HDR with tone-mapped SDR versions, including expert grades; BrightVQ (Saini et al., 2026b) and Beyond8Bits (Saini et al., 2026a) extend subjective HDR-UGC quality data to larger scales and inform HDR-aware quality models for converted content. Our corpus adds an open Netflix Open Content master (Netflix, 2026). Luma-Eval fixes one run-by-us protocol across all methods; redistributable content ships directly, while restricted portions are represented by deterministic scripts, content IDs, manifests, and checksums.