LumaGuide: Distribution Shaping for Training-Free HDR Generation in Diffusion Models

Bowen Chen[1], **Shreshth Saini**[1], **Balu Adsumilli**[2], **Alan C. Bovik**[1]

[1]The University of Texas at Austin, [2]Google, Inc.

Pretrained diffusion models generate realistic images, but their outputs remain constrained by the statistical biases of their training data, limiting their ability to produce high dynamic range (HDR) content. In this work, we introduce **LumaGuide**, a training-free framework for distribution shaping in diffusion models. Instead of modifying model parameters, **LumaGuide** steers the sampling process to match target feature distributions via differentiable energy-based guidance. We instantiate this framework for HDR generation by controlling luminance distributions in perceptually uniform PQ space. Our results show that aligning luminance histograms can induce HDR-consistent behavior, including coherent highlights and preserved shadow detail, while maintaining semantic fidelity. Beyond HDR, **LumaGuide** enables flexible specification of target distributions through data-driven presets, reference images, or text-driven predictors, and extends naturally to video generation with temporal consistency constraints. More broadly, our work demonstrates that controllable generation can be achieved by directly shaping output distributions at sampling time, without retraining diffusion models.

Date: May 13, 2026
Page: https://shreshthsaini.github.io/LumaGuide
Correspondence: bwchen@utexas.edu, saini.2@utexas.edu

1 Introduction

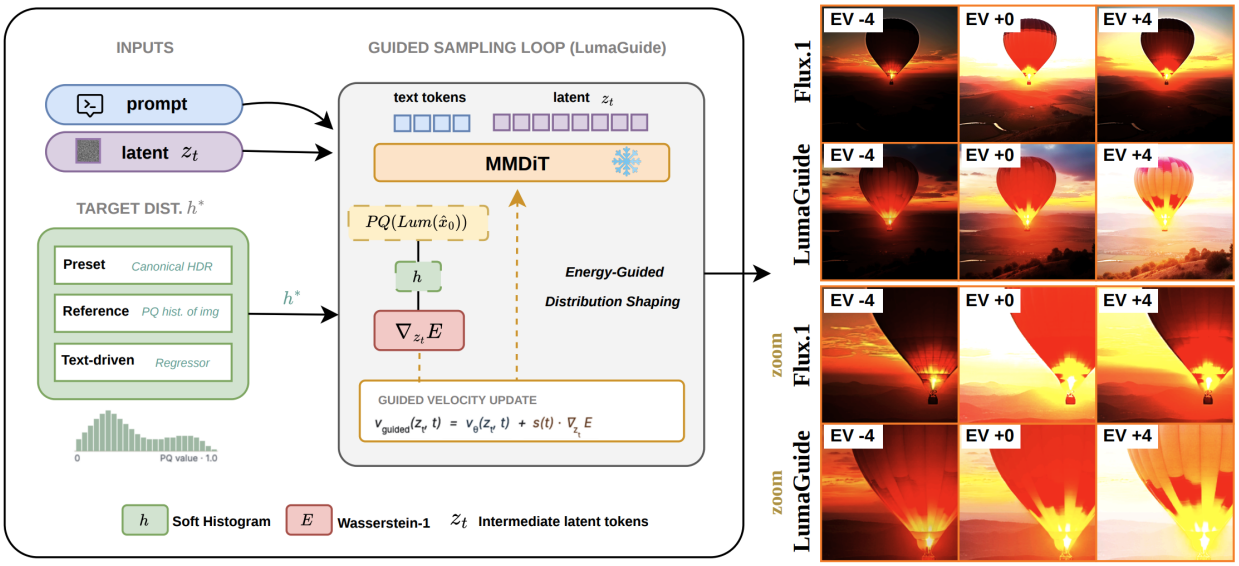


Figure 1 **LumaGuide** steers a pretrained diffusion model toward HDR-consistent luminance distributions at sampling time, without modifying any model weights. Generated outputs preserve semantic content (top row) while producing structured highlights and faithful shadow detail under exposure adjustment (bottom row), matching target HDR statistics in PQ space.

Pretrained diffusion models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Saharia et al., 2022;

Ramesh et al., 2022) have demonstrated remarkable abilities to generate realistic images from text prompts, yet their outputs remain implicitly constrained by the statistical properties of their training data. In particular, models trained on internet-scale corpora of images that are typically quantized and compressed for legacy 8-bit displays internalize a Standard Dynamic Range (SDR) prior. This limits their ability to represent the inherently wide dynamic range of natural light, which spans several orders of magnitude in luminance (Reinhard et al., 2010; International Telecommunication Union, 2025). As a result, even highly realistic text-to-image diffusion models often fail to synthesize physically plausible high-intensity highlights or coherent HDR luminance structure. Existing approaches to HDR generation typically rely on additional reconstruction pipelines or modifications to the generative model itself. Prior work has explored multi-stage reconstruction or exposure expansion pipelines (Eilertsen et al., 2017; Marnerides et al., 2018, 2021), while more recent diffusion-based approaches operate in perceptually uniform domains such as PQ or PU21 (International Telecommunication Union, 2025; Azimi et al., 2021). For example, X2HDR (Wu et al., 2026) shows that pretrained VAEs can faithfully encode HDR signals in perceptually uniform space, but still relies on fine-tuning the diffusion backbone, e.g., via LoRA (Hu et al., 2022), to overcome SDR-biased generation. Other diffusion-based HDR methods synthesize exposure brackets or fuse latent exposures to recover HDR content (Bemana et al., 2025; Wang et al., 2025; Yu et al., 2026). These approaches are effective, but they often introduce additional training, inference, or architectural complexity, reducing flexibility across backbones and tasks.

In this work, we take a different perspective. Instead of modifying model parameters, we directly control the output statistics of a pretrained diffusion model at sampling time. We formalize this as distribution shaping, in which sampling is steered toward target output statistics via differentiable energy guidance. This perspective relates to guided sampling methods in diffusion models, such as classifier- and classifier-free guidance (Dhariwal and Nichol, 2021; Ho and Salimans, 2022), as well as broader training-free guidance methods that steer diffusion models using arbitrary differentiable objectives or energy functions (Bansal et al., 2023; Chung et al., 2022; Yu et al., 2023; Park et al., 2023). However, our method differs in that it shapes low-level perceptual distributions rather than semantic, class-conditional, or task-specific reconstruction signals. We instantiate this framework for text-to-HDR image generation. Given a text prompt, our goal is to synthesize an image whose luminance statistics are consistent with HDR appearance while preserving the semantic content and spatial coherence produced by the pretrained diffusion prior. We observe that luminance distributions in perceptually uniform PQ space capture a key aspect of HDR appearance (International Telecommunication Union, 2025; Mantiuk et al., 2011; Azimi et al., 2021). This motivates luminance histogram alignment as a natural objective for encouraging HDR-consistent rendering.

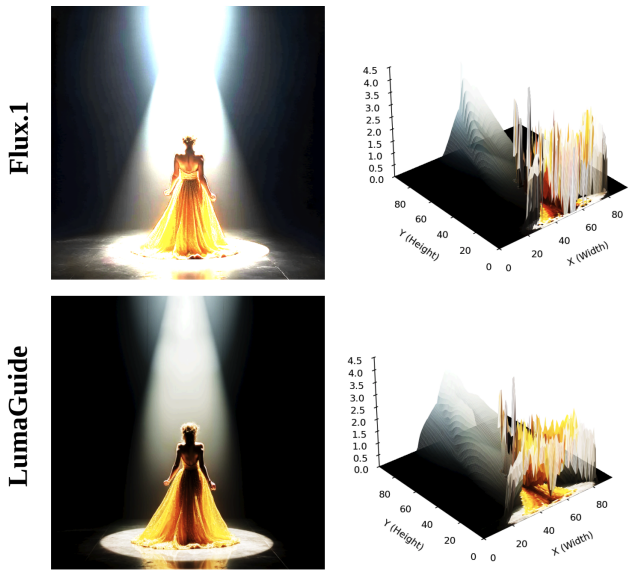


Figure 2 Per-pixel PQ-luminance surface for an identical seed and prompt. The Flux.1 (Black Forest Labs, 2024) baseline (top) overshoots into clipped highlights; LumaGuide (bottom) redistributes mass into mid-tones and preserves the highlight gradient, matching the target HDR distribution.

To this end, we introduce **LumaGuide**, a training-free framework for HDR distribution shaping in diffusion models. During sampling, **LumaGuide** computes a differentiable soft histogram of the predicted image in PQ space and minimizes a Wasserstein-1 distance (W_1) (Arjovsky et al., 2017) to the target distribution. The resulting gradient is backpropagated through the VAE to steer the diffusion trajectory in latent space. Since the histogram constraint is permutation-invariant, it does not explicitly impose spatial structure, allowing the diffusion prior to preserve geometry and semantics while adjusting global luminance statistics. This decoupling of semantic generation and distribution control also makes LumaGuide naturally applicable to video generation. By applying the same distribution shaping principle to pretrained video diffusion models (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2024), we enable zero-shot HDR video synthesis. To

maintain temporal consistency under distribution shaping, we introduce a Temporal Luminance Coherence (TLC) term that penalizes highlight flickering across frames, inspired by prior work on temporal consistency in video synthesis (Lai et al., 2018). This results in a training-free framework for controllable HDR video generation that maintains motion and structural fidelity. In summary, our contributions are as follows:

- **Distribution Shaping Framework**: We propose a training-free framework for controlling output feature distributions in diffusion models via differentiable energy guidance at sampling time.
- **LumaGuide for Text-to-HDR Generation**: We instantiate this framework for HDR image synthesis, showing that luminance histogram alignment in perceptually uniform space provides an effective mechanism for inducing HDR-consistent outputs. We further extend the framework to video generation, introducing a TLC term to improve temporal stability without additional training.
- **Theoretical Guarantees**: We show that energy-guided distribution shaping admits controlled energy descent (Theorem 1), that permutation-invariant feature constraints do not explicitly impose spatial rearrangements (Proposition 1), and that global affine brightness adjustments cannot generally reproduce target HDR luminance distributions (Theorem 2).
- **Empirical Validation**: Extensive experiments across image and video backbones (e.g., Flux.1 (Black Forest Labs, 2024), SD3 (Esser et al., 2024), SDXL (Podell et al., 2023), and CogVideoX (Yang et al., 2024)) demonstrate significant improvements in distribution alignment and HDR fidelity over existing methods.

2 Distribution Shaping in Diffusion Models

We formalize *distribution shaping* as the problem of steering a pretrained diffusion model so that a designated low-dimensional feature of its output matches a prescribed target distribution, without any modification to model weights. We then establish that this objective admits a principled energy-guided solution with provable guarantees on energy descent and on preservation of spatial structure.

2.1 Problem Formulation

Let p_θ denote the implicit distribution of a pretrained flow-matching model with velocity field $v_\theta : \mathcal{Z} \times [0,1] \to \mathcal{Z}$ and decoder $\mathrm{Dec} : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X} = \mathbb{R}^{C \times H \times W}$. Let $f : \mathcal{X} \to \mathbb{R}^N$ be a *pixel-wise feature map*, i.e. $f(x)_i = \varphi(x_{\cdot,i})$ for some scalar map φ acting on the channel vector at spatial location $i \in \{1, \dots, N\}$, $N = H \cdot W$. The *empirical feature distribution* of $x \in \mathcal{X}$ under f is the random measure (Empirical feature distribution):

$$P_f(x) \;=\; \tfrac{1}{N} \sum_{i=1}^{N} \delta_{f(x)_i} \;\in\; \mathcal{P}(\mathbb{R}). \tag{1}$$

Given a target $P^* \in \mathcal{P}(\mathbb{R})$ and a metric $\mathcal{D}$ on $\mathcal{P}(\mathbb{R})$, distribution shaping seeks to reduce

$$\mathcal{E}(z) \;:=\; \mathcal{D}\big(P_f(\mathrm{Dec}(z)),\, P^*\big) \tag{2}$$

along the sampling trajectory induced by the pretrained model, without modifying v_θ or Dec.

Equation (2) differs sharply from posterior sampling or classifier guidance (Chung et al., 2022; Ho and Salimans, 2022; Saini et al., 2025b): the constraint is imposed on a statistic of the output, rather than on its identity. This distinction allows us to retain the pretrained prior as the source of semantic and spatial structure while shaping marginal output behavior. In practice, $P_f(x)$ is approximated using a differentiable soft histogram.

2.2 Energy-Guided Sampling

We adopt a soft histogram approximation $h_\sigma(x) \in \mathbb{R}^K$ of $P_f(x)$, with bin centers $\{b_k\}_{k=1}^K$ and Gaussian kernel of bandwidth σ:

$$[h_\sigma(x)]_k \;=\; \frac{1}{N} \sum_{i=1}^{N} \exp\!\left(-\frac{(f(x)_i - b_k)^2}{2\sigma^2}\right) / Z_i, \qquad Z_i = \sum_{k'} \exp\!\left(-\tfrac{(f(x)_i - b_{k'})^2}{2\sigma^2}\right). \tag{3}$$

The energy used in our framework is

$$E(x) \;=\; \mathcal{D}\big(h_\sigma(x),\, h^*\big), \tag{4}$$

where h^* is the target histogram and $\mathcal{D}$ is the Wasserstein-1 distance (Arjovsky et al., 2017) on K-bin distributions (Section 3). Because the clean output is unavailable during sampling, we evaluate E on the flow estimate (Chung et al., 2022; Efron, 2011; Kim and Ye, 2021) $\hat{x}_0(z_t) := \mathrm{Dec}\big(z_t - t\, v_\theta(z_t, t)\big)$. The guided velocity is

$$v_{\mathrm{g}}(z_t, t) \;=\; v_\theta(z_t, t) \;+\; s(t)\, \nabla_{z_t} E(\hat{x}_0(z_t)), \tag{5}$$

with time-dependent guidance scale $s : [0, 1] \to \mathbb{R}_{\geq 0}$. Since the sampler steps as $z_{t_{n+1}} = z_{t_n} + (t_{n+1} - t_n) v_{\mathrm{g}}$ with $t_{n+1} < t_n$, the added velocity term induces a negative-gradient step on E.

The next result establishes that this update is a descent method on E in expectation, with a quantifiable noise term controlled by the diffusion model's own velocity error.

Theorem 1 (Energy descent). *Assume $z \mapsto E(\hat{x}_0(z))$ is L-smooth and the velocity error has bounded second moment, $\mathbb{E}\|v_\theta - v^*\|^2 \leq \sigma_v^2$. Let $\eta_t := |t_{n+1} - t_n| s(t)$ be the effective guidance step size. If $\eta_t \in (0, 2/L]$, then*

$$\mathbb{E}\big[\Delta E(z_t)\big] \;\leq\; -\eta_t \Big(1 - \tfrac{L\,\eta_t}{2}\Big)\, \mathbb{E}\|\nabla_z E\|^2 \;+\; C\, \sigma_v^2\, t^2\, \Delta t, \tag{6}$$

where $\Delta E(z_t) := E(\hat{x}_0(z_{t+\Delta t})) - E(\hat{x}_0(z_t))$ and C is a constant depending only on the decoder. The deterministic term is maximized at $\eta_t = 1/L$.

The first (negative) term is the guidance benefit, proportional to the squared gradient. The second (positive) term is the noise penalty, growing with the velocity-prediction error σ_v^2 and the timestep t. The energy decreases on average whenever the gradient signal exceeds the noise floor set by $\sigma_v^2 t^2$. Full proof in Appendix B.6.

2.3 Why Distribution Shaping Preserves Semantics

The most striking empirical phenomenon in LumaGuide is that, despite applying a strong global energy gradient, semantic content and geometric layout remain largely intact. We now formalize the structural reason for this: such gradients do not explicitly encode spatial rearrangement, but act through local feature values and global histogram statistics.

Proposition 1 (Spatial decoupling). *Let $E(x) = \mathcal{D}(h_\sigma(x), h^*)$ for a pixel-wise feature $f(x)_i = \varphi(x_{\cdot,i})$. Then:*

(i) ***Permutation equivariance.*** *For any spatial permutation $\pi \in S_N$, $E(\pi \cdot x) = E(x)$ and $\nabla_x E(\pi \cdot x) = \pi \cdot \nabla_x E(x)$.*

(ii) ***Pixel-wise gradient form.*** *The gradient at site i depends only on the feature value $f(x)_i$ and on the global histogram $h_\sigma(x)$:*

$$\big[\nabla_x E(x)\big]_{\cdot,i} \;=\; \nabla_{x_{\cdot,i}} \varphi(x_{\cdot,i}) \cdot g\big(\varphi(x_{\cdot,i});\, h_\sigma(x), h^*\big), \tag{7}$$

for a scalar function g.

(iii) ***No spatial ordering information.*** *E depends on x only through the multiset $\{f(x)_i\}_{i=1}^N$, and therefore cannot distinguish images that differ only by a spatial permutation.*

See Appendix B.7 for proof. Algo. 1 shows the full algorithm. It requires only a forward pass, the standard flow estimate, one backward pass through the decoder per step, and no additional model training.

3 LumaGuide for HDR Generation

We instantiate the framework using three deliberate design choices: PQ encoding, Wasserstein-1 distance, and a constant guidance schedule, supported by ablations in Section 4 and the analysis below.

Algorithm 1 LumaGuide: Energy-Guided Distribution Shaping

Require: Pretrained flow model v_θ, VAE decoder Dec, target histogram $h^* \in \Delta^{K-1}$, schedule $s(t)$, discretization $\{t_n\}_{n=0}^{T}$ with $t_{n+1} < t_n$, soft-histogram bandwidth σ.

Sample $z_{t_0} \sim \mathcal{N}(0, I)$
for $n = 0, \ldots, T-1$ **do**
  $\hat{x}_0 \leftarrow \mathrm{Dec}\big(z_{t_n} - t_n\, v_\theta(z_{t_n}, t_n)\big)$ ▷ flow decode ($\hat{x}_0$ is RGB)
  $Y \leftarrow \mathrm{PQ}(\mathrm{Lum}(\hat{x}_0))$ ▷ Per-pixel PQ luminance
  $h_\sigma \leftarrow$ soft-histogram of Y via Eq. (3)
  $E \leftarrow W_1(h_\sigma, h^*)$ ▷ Wasserstein-1 distance, perceptually weighted
  $g_t \leftarrow \nabla_{z_{t_n}} E$ ▷ Backprop through Dec and PQ $\circ$ Lum
  $z_{t_{n+1}} \leftarrow z_{t_n} + (t_{n+1} - t_n)\big(v_\theta(z_{t_n}, t_n) + s(t_n)\, g_t\big)$
return $\hat{x}_0(z_{t_T})$

3.1 PQ-Space Feature Map

We take $\varphi = \mathrm{PQ} \circ \mathrm{Lum}$ where $\mathrm{Lum}(x) = 0.2627R + 0.6780G + 0.0593B$ is BT.2020 luminance and $\mathrm{PQ} : \mathbb{R}_{\geq 0} \to [0, 1]$ is the SMPTE ST.2084 OETF.

Proposition 2 (PQ-space gradient stability)**.** *Let $L \in [L_{\min}, L_{\max}]$ denote scene luminance in nits with $L_{\max}/L_{\min} \geq 10^4$ (HDR range). For a fixed bin grid in PQ space,*

(i) PQ *is monotone and smooth on $[L_{\min}, L_{\max}]$, and $\mathrm{PQ}'(L)L$ is bounded above and below by positive constants. Hence per-pixel gradients converted through PQ are scale-balanced across the dynamic range.*

(ii) *In linear space, the inverse PQ Jacobian induces polynomially mismatched scales across luminance ranges; for large L, the corresponding factor scales as $\Theta(L^{1-1/m_2})$, yielding gradient variance that grows as $\Omega(L_{\max}^{2(1-1/m_2)})$.*

(iii) *Consequently, the noise-to-signal ratio of the histogram-shaping gradient in PQ space is bounded independently of $L_{\max}$, while in linear space it grows polynomially with the dynamic range.*

See Appendix B.8 for proof. This proposition explains the dominant effect of domain choice in Table 3: linear-space gradients exhibit scale mismatches across luminance ranges, making stable step-size selection difficult.

3.2 Wasserstein-1 over KL and ℓ_2

The choice of distribution metric is important. HDR luminance histograms in PQ space are defined by a sparse high-percentile tail; a small fraction of pixels carry the highlights that distinguish HDR from SDR, and the optimization signal must reach those bins. Standard distributional metrics differ sharply in whether they can. On the real line, the Wasserstein-1 distance admits the closed-form CDF representation $W_1(h, h^*) = \sum_{k=1}^{K-1}(b_{k+1} - b_k)|F_k - F_k^*|$ (Arjovsky et al., 2017), where F, F^* are the discrete CDFs of h, h^* on the ordered grid $\{b_1 < \cdots < b_K\}$. The subgradient $\partial W_1 / \partial h_k$ is therefore bounded and remains non-zero whenever the relevant CDF gaps are non-zero. Crucially, its magnitude depends on the bin spacing and the CDF gap, *not* on the bin probabilities themselves, so even when $h_k \ll 1$ in a sparse highlight bin, the gradient driving mass toward that bin remains well-conditioned.

KL and ℓ_2 fail in this regime, in opposite directions. The KL gradient $\partial \mathrm{KL}(h \| h^*)/\partial h_k = \log(h_k / h_k^*) + 1$ can become unbounded near sparse bins, producing unstable updates where reliable tail statistics are most needed. The ℓ_2 gradient $2(h_k - h_k^*)$ depends only on pointwise bin differences and does not exploit the ordering of luminance bins, making it less effective at transporting mass toward sparse highlight regions. Neither metric respects the ordering of luminance values; distant bins are treated as exchangeable rather than as a transport problem on the line. Table 3 bears this out empirically.

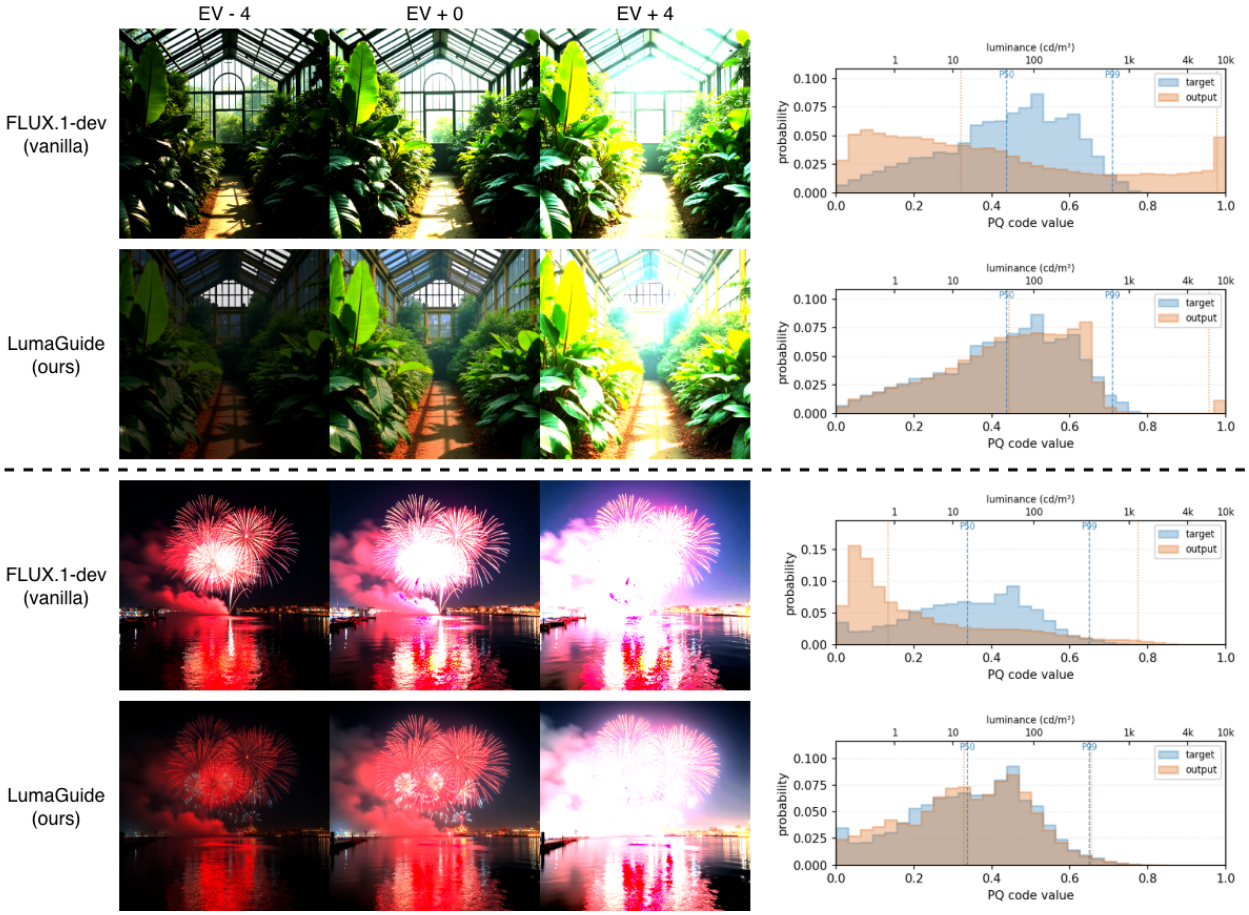


Figure 3 HDR visualization under different exposure levels. In PQ space, Flux.1 (Black Forest Labs, 2024) often yields over-exposed, non-HDR luminance distributions. **LumaGuide** steers the luminance distribution toward a target HDR distribution while preserving scene structure.

3.3 Guidance Schedule

Following previous methods (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Yu et al., 2023; Black Forest Labs, 2024), we adopt a constant guidance schedule $s(t) = s_0$ throughout sampling. We confirm this in our setting (Table 8), where the constant schedule outperforms an SNR-weighted bell-shaped schedule and a time-windowed variant on every reported metric. Appendix D discusses this in more detail.

3.4 Extension to Video Generation

The same distribution shaping principle extends to video by enforcing luminance constraints at each frame, but naive frame-wise guidance may introduce temporal inconsistencies, particularly in high-luminance regions. To reduce this effect, we introduce a Temporal Luminance Coherence (TLC) term that penalizes abrupt fluctuations in these regions. Let h_τ be the soft histogram of frame τ, and let Ω_τ be a high-luminance mask. We define

$$\mathcal{L}_{\text{TLC}} = \sum_{\tau=2}^{T} \left\| (Y_\tau - Y_{\tau-1}) \odot \Omega_\tau \right\|_2^2, \qquad E_{\text{video}} = \sum_{\tau=1}^{T} W_1(h_\tau, h^*) + \lambda_{\text{TLC}} \mathcal{L}_{\text{TLC}}. \tag{8}$$

Uniform target specification: Distribution shaping admits three target-specification modes $h^* \in \Delta^{K-1}$ within a single objective: (a) preset distributions, h^* fixed offline; (b) reference-based, $h^* = h_\sigma(x_{\text{ref}})$; (c) text-driven, $h^* = \text{MLP} \circ \text{CLIP}(c)$ for caption c. All three plug into Algorithm 1 without altering sampling. Since Δ^{K-1} is convex, their mixtures remain valid targets under the same objective.

4 Experiments

Our primary experiments are conducted using Flux.1-dev (Black Forest Labs, 2024), SD3 (Esser et al., 2024), and SDXL (Podell et al., 2023), without modifying model parameters or architectures. For video generation, we extend the same distribution shaping principle to CogVideoX (Yang et al., 2024). More detailed results are provided in Appendices F and I. **LumaGuide** is plug-and-play for diffusion-based image and video models.

4.1 Experimental Setup

Unless otherwise specified, all experiments are performed at a resolution of 512×512. We use 28 sampling steps and a classifier-free guidance (CFG) scale of 3.5. All evaluations are conducted on a set of 100 HDR-oriented prompts covering diverse luminance conditions. Luminance distributions are represented using $K = 32$ histogram bins in PQ space. We employ differentiable soft binning using a Gaussian kernel with standard deviation $\sigma = 0.5/K$. Our default configuration uses a base guidance scale of $s_0 = 2000$, together with a perceptual-log Wasserstein-1 objective, a constant guidance schedule, and integral normalization. By default, target luminance distributions are obtained using the text-driven regressor described in Appendix G.

For quantitative evaluation, we follow (Wu et al., 2026) and perform all comparisons in PQ-encoded space using Q-Eval (Zhang et al., 2025) as a proxy for measuring perceptual quality and alignment for HDR image generation. We use uW_1 for unweighted Wasserstein-1 distance between generated and target PQ-luminance histograms over $K = 32$ bins, $p50_{\text{dist}}$ and $p99_{\text{dist}}$ for absolute percentile errors in PQ space, and nits_{99} for the 99th-percentile luminance after inverse PQ conversion. DR_{stops} reports the robust luminance span in exposure stops, while JOD is the Bradley–Terry just-objectionable-difference score from the subjective study. We further discuss the details of HDR evaluation metrics in Appendix J. All experiments are conducted on a single NVIDIA A100 GPU (40GB). For high-resolution generation (1024×1024), we employ VAE gradient checkpointing and spatial tiling to enable memory-efficient backpropagation during guided sampling. Video experiments are conducted on a single NVIDIA H200 GPU.

4.2 Distribution Control and HDR Emergence

We first evaluate whether **LumaGuide** can effectively control luminance distributions and induce HDR-consistent rendering behavior. As shown in Figure 3, the proposed guidance consistently reshapes the generated PQ luminance histograms toward the target HDR distribution while preserving semantic structure and spatial coherence.

Increasing the guidance scale generally improves distribution alignment, with the Wasserstein-1 distance decreasing substantially up to $s_0 = 3000$ before slightly degrading at $s_0 = 5000$ (Table 1). Importantly, the guidance does not simply increase overall brightness. Instead, it redistributes luminance mass from over-saturated highlight regions toward mid-tones, resulting in more balanced luminance allocation and better preservation of highlight structure. This effect is particularly visible under exposure adjustment in Figure 3, where baseline generations often exhibit clipped highlights and unstable luminance behavior when interpreted in PQ space, while **LumaGuide** produces more coherent highlight roll-off and preserved shadow detail characteristic of HDR content.

Table 1 Effect of guidance scale s_0 on distribution alignment and luminance statistics. We select $s_0 = 2000$ as the default trade-off point.

s_0	uW1 ↓	$p50_{\text{dist}}$ ↓	$p99_{\text{dist}}$ ↓	nits_{99}
0	3.79	0.116	0.207	4867
10	3.77	0.115	0.206	4861
100	3.42	0.104	0.198	4658
500	2.16	0.071	0.153	3558
1000	1.25	0.045	0.105	2648
2000	0.58	0.024	0.053	1694
3000	0.49	0.020	0.041	1425
5000	0.52	0.019	0.045	1496

At excessively large guidance scales, the optimization may become over-constrained, leading to artifacts such as banding and unnatural textures. Additional analysis of these failure cases is provided in Appendix E.2. We select $s_0 = 2000$ as the default trade-off point.

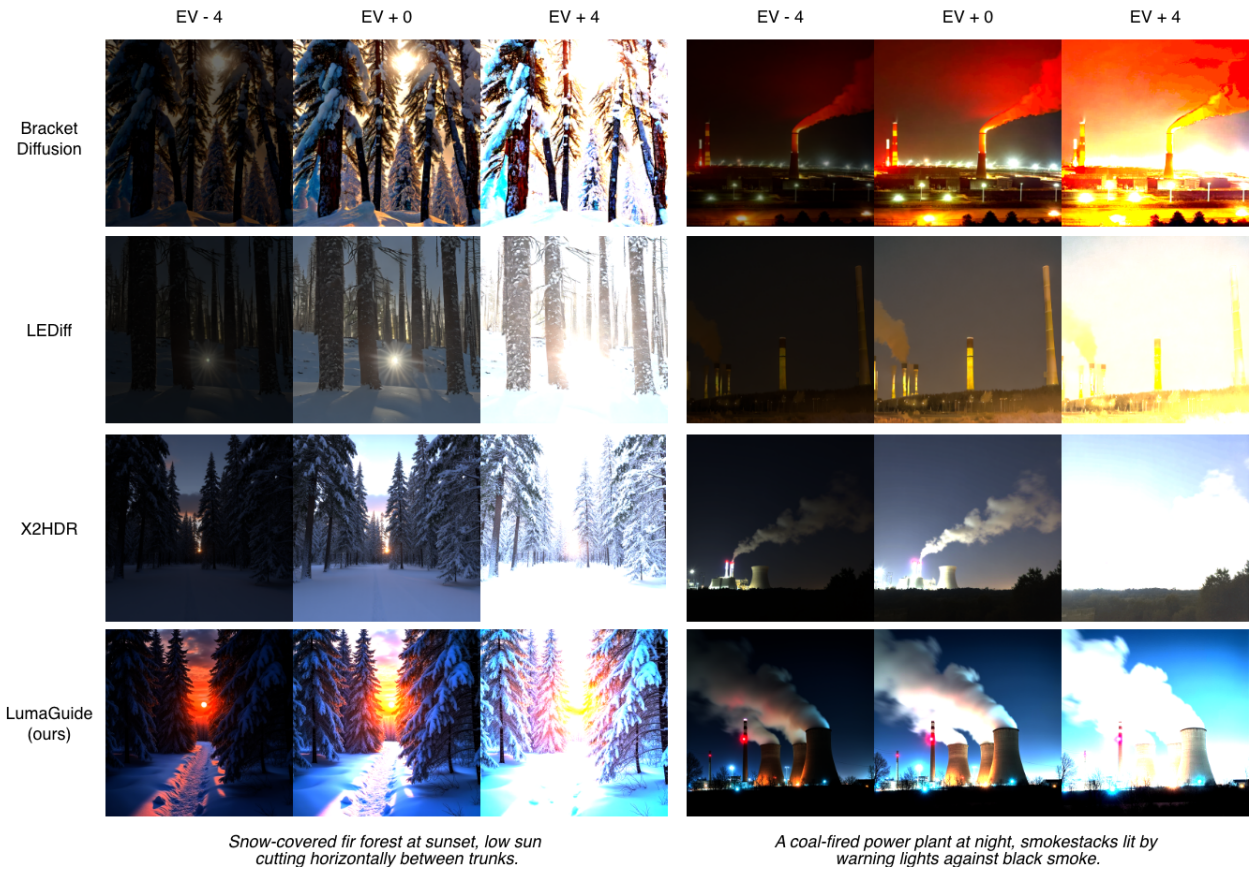


Figure 4 Qualitative comparison with existing HDR generation methods. For fair comparison, we normalize the exposure of all methods by adjusting the EV+0 images such that the median luminance is fixed at 8 nits since several existing methods are not calibrated to absolute luminance values, making direct visual comparison otherwise unreliable.

4.3 Comparisons with Existing Models

Table 2 compares **LumaGuide** with existing HDR generation methods. Although X2HDR (Wu et al., 2026) achieves slightly higher Q-quality, it relies on fine-tuning the diffusion backbone, whereas **LumaGuide** remains entirely training-free and preserves the pretrained model unchanged. Despite this, our method achieves the best Q-alignment score and the largest dynamic range among all compared approaches, indicating stronger HDR characteristics and more faithful luminance allocation.

These improvements are also reflected perceptually. Figure 5 shows that **LumaGuide** achieves the highest preference in our JOD-based subjective study, outperforming all baselines in overall HDR quality. Qualitative comparisons in Figure 4 further demonstrate improved highlight structure and shadow preservation under different exposure settings.

In addition, **LumaGuide** remains computationally efficient, introducing only moderate overhead over vanilla Flux and runtime comparable to X2HDR (Wu et al., 2026), while remaining substantially faster than BracketDiffusion (Bemana et al., 2025).

Table 3 Ablation of feature domain and distribution distance. PQ-space guidance significantly improves distribution alignment over linear-domain variants. Within PQ space, Wasserstein-1 (W_1) further outperforms ℓ_2 and KL by enabling ordered transport across histogram bins.

Setting	Domain	Distance	uW1 ↓	$p50_{\text{dist}}$ ↓	$p99_{\text{dist}}$ ↓	DR_{stops} ↑
Linear + W_1	Linear	W_1	3.73	0.115	0.199	16.37
Linear + ℓ_2	Linear	ℓ_2	3.79	0.116	0.207	16.33
PQ + ℓ_2	PQ	ℓ_2	3.40	0.100	0.206	16.18
PQ + KL	PQ	KL	2.06	0.065	0.143	**16.51**
PQ + W_1	PQ	W_1	**0.58**	**0.024**	**0.053**	14.99

Table 2 Comparison with HDR generation baselines. LumaGuide achieves the best alignment and dynamic range with competitive quality and moderate runtime.

Method	Q-quality ↑	Q-alignment ↑	DR_{stops} ↑	JOD ↑	Time ↓
LEDiff (Wang et al., 2025)	0.425	0.612	4.71	-0.88	∼8.6 s
BracketDiffusion (Bemana et al., 2025)	0.448	0.648	12.25	-0.30	∼389 s
X2HDR (Wu et al., 2026)	**0.579**	0.773	11.41	+0.43	∼**6 s**
LumaGuide	0.568	**0.814**	**14.99**	**+0.75**	7.8 s

4.4 Analysis and Ablations

We analyze key design choices in **LumaGuide** to understand whether the observed improvements arise from principled distribution shaping rather than trivial transformations. More detailed ablation studies can be found in Appendix E.

Effect of domain and distance. Table 3 studies the impact of feature domain (linear vs. PQ) and distance metric (W_1, ℓ_2, KL). Linear-domain guidance performs similarly to the unguided baseline, indicating that linear luminance does not provide perceptually meaningful gradients for distribution shaping. In contrast, PQ-space guidance consistently improves alignment, even with simple ℓ_2 distance, highlighting the importance of perceptual reparameterization. Within PQ space, W_1 further outperforms both ℓ_2 and KL divergence, likely because it preserves the ordering of histogram bins and enables smoother luminance transport across neighboring bins.

Highlight-aware distribution shaping. Table 5 compares different histogram weighting strategies for highlight control. Manual weighting provides moderate alignment but limited perceptual quality. Dual-histogram weighting aggressively emphasizes highlight bins but introduces noticeable degradation in distribution alignment. In contrast, the proposed perceptual-log weighting achieves the best balance, producing the lowest $p99_{\text{dist}}$ among weighting strategies, together with the highest perceptual quality and alignment scores. This suggests that logarithmic damping stabilizes optimization while preserving sensitivity to high-luminance regions.

4.5 Flexible Control and Video Extension

A key advantage of **LumaGuide** is that the target distribution $\mathcal{P}^*$ can be specified externally. We support three modes: (1) user-driven presets, where expert users directly define target luminance histograms over PQ bins; (2) reference-based specification, where the PQ luminance histogram is extracted from a reference image; and (3) text-driven specification, where a lightweight regressor maps text descriptions generated by Qwen2.5-7B-Instruct (Team, 2024) to luminance histograms derived from the Beyond8Bits dataset (Saini et al., 2026a, 2025a, 2026b). The regressor is used solely for target prediction and does not modify the diffusion backbone; implementation details are provided in Appendix G.

Table 4 Cross-backbone comparison of LumaGuide at the selected operating points for each model.

Model	Q-quality ↑	Q-alignment ↑	DR (stops) ↑
Flux	**0.568**	**0.814**	14.99
SD3	0.512	0.795	15.42
SDXL	0.431	0.655	**15.90**

Importantly, **LumaGuide** is model-agnostic and can be applied to pretrained diffusion models without retrain-

Metric	Manual	Dual-hist.	Perc.-log
uW1 ↓	0.70	1.99	**0.58**
$p99_{\text{dist}}$ ↓	0.055	0.206	**0.053**
DR_{stops} ↑	14.38	**16.84**	14.99
Q-quality ↑	0.514	0.325	**0.568**
Q-alignment ↑	0.708	0.299	**0.814**

Table 5 Ablation of histogram bin weighting strategies. Perceptual-log weighting gives the best results.

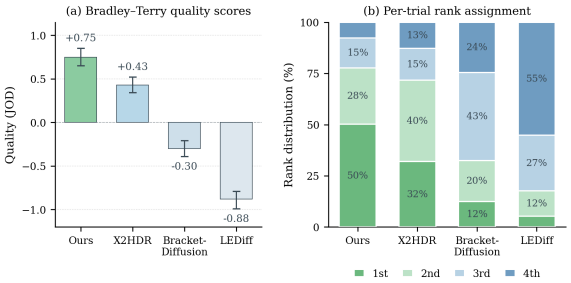


Figure 5 Subjective study. Left: Bradley–Terry JOD scores with 95% CIs. **Right:** Per-trial rank distribution. LumaGuide is ranked first in 50% of trials.

ing. Table 4 shows results on Flux.1 (Black Forest Labs, 2024), SD3 (Esser et al., 2024), and SDXL (Podell et al., 2023), demonstrating consistent luminance distribution control and HDR characteristics across backbones. We further extend **LumaGuide** to video generation by applying the same distribution shaping objective independently to each frame. To mitigate temporal luminance instability, we incorporate the Temporal Luminance Coherence (TLC) term introduced in Section 3.4. Additional image and video results are provided in Appendices F and I.

5 Conclusion

We presented LumaGuide, a training-free framework for distribution shaping in diffusion models. By steering the sampling process toward target feature distributions, our method enables direct control over output statistics without modifying model parameters. Instantiated for HDR generation, LumaGuide shows that shaping luminance distributions can induce HDR-consistent behavior while preserving semantic and spatial fidelity. More broadly, our results suggest that controllable generation can be achieved by directly shaping output distributions at sampling time. Extending this framework to richer spatially structured or multi-modal feature distributions remains an important direction for future work.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

Maryam Azimi et al. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 843–852, 2023.

Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Bracket diffusion: Hdr image generation by consistent ldr denoising. In *Computer Graphics Forum*, volume 44, page e70086. Wiley Online Library, 2025.

Black Forest Labs. FLUX.1: Text-to-image generation models. https://github.com/black-forest-labs/flux, 2024.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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

International Telecommunication Union. Recommendation ITU-R BT.2100-3: Image parameter values for high dynamic range television for use in production and international programme exchange. https://www.itu.int/rec/R-REC-BT.2100, February 2025.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Kwanyoung Kim and Jong Chul Ye. Noise2Score: Tweedie's approach to self-supervised image denoising without clean images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.

Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer graphics forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

Demetris Marnerides, Thomas Bashford-Rogers, and Kurt Debattista. Deep hdr hallucination for inverse tone mapping. *Sensors*, 21(12):4032, 2021.

Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:76382–76408, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Shreshth Saini, Alan C. Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. Chug: Crowdsourced user-generated hdr video quality dataset. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2504–2509, 2025a. doi: 10.1109/ICIP55913.2025.11084488.

Shreshth Saini, Shashank Gupta, and Alan C. Bovik. Rectified-cfg++ for flow based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Shreshth Saini, Bowen Chen, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Seeing beyond8bits: Subjective and objective quality assessment of hdr-ugc videos. *arXiv preprint arXiv:2603.00938*, 2026a.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang. Lediff: Latent exposure diffusion for hdr generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 453–464, 2025.

Ronghuan Wu, Wanchao Su, Kede Ma, Jing Liao, and Rafał K Mantiuk. X2hdr: Hdr image generation in a perceptually uniform space. *arXiv preprint arXiv:2602.04814*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

Zhengming Yu, Li Ma, Mingming He, Leo Isikdogan, Yuancheng Xu, Dmitriy Smirnov, Pablo Salamanca, Dao Mi, Pablo Delgado, Ning Yu, et al. Diffhdr: Re-exposing ldr videos with video diffusion models. *arXiv preprint arXiv:2604.06161*, 2026.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, et al. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10621–10631, 2025.

Appendix

Contents

A Related Work

A.1 HDR Generation Methods

Image generation has seen rapid progress with deep generative models. Early approaches relied on GAN-based methods (Goodfellow et al., 2020; Karras et al., 2020) and autoregressive models (Ramesh et al., 2021; Yu et al., 2022) to synthesize realistic images, but these models often struggled with training stability, likelihood-quality trade-offs, or scalability. More recently, diffusion models (Ho et al., 2020; Song et al., 2020) have emerged as the dominant paradigm for high-fidelity image synthesis, achieving strong performance in class-conditional and text-to-image generation (Dhariwal and Nichol, 2021; Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022).

Extending these models to HDR generation presents additional challenges, as pretrained diffusion models are inherently biased toward SDR data distributions. Prior work on HDR reconstruction and inverse tone mapping expands dynamic range from SDR inputs using supervised learning, saturated-region reconstruction, or exposure-based techniques (Eilertsen et al., 2017; Marnerides et al., 2018, 2021).

More recent work has explored adapting diffusion models for HDR synthesis. Bracket Diffusion (Bemana et al., 2025) synchronizes multiple LDR exposure brackets from pretrained LDR diffusion models to reconstruct HDR images without retraining. LEDiff (Wang et al., 2025) introduces latent exposure fusion and trains HDR-specific components to recover highlight and shadow details. X2HDR (Wu et al., 2026) demonstrates that pretrained VAEs can encode HDR signals in perceptually uniform domains such as PQ or PU21 (International Telecommunication Union, 2025; Azimi et al., 2021), but still requires fine-tuning the diffusion backbone, e.g., via LoRA (Hu et al., 2022), to mitigate SDR bias. Other recent efforts further explore diffusion-based LDR-to-HDR reconstruction for videos (Yu et al., 2026).

In contrast, our approach does not modify model parameters. Instead, we directly control the output luminance distribution of a pretrained diffusion model at sampling time, enabling HDR-consistent generation without retraining.

A.2 Classifier Guidance and Energy-Based Steering

Controlling diffusion models at inference time has been widely studied through guidance techniques. Early methods introduce classifier guidance, where gradients from an external classifier are used to steer the sampling process toward desired classes (Dhariwal and Nichol, 2021). Classifier-free guidance further simplifies this approach by jointly learning conditional and unconditional scores, enabling efficient trade-offs between fidelity and diversity without an external classifier (Ho and Salimans, 2022).

Beyond class-conditional control, recent work has generalized guidance to broader differentiable objectives. Score-based generative modeling provides a natural framework for incorporating conditional gradients into the reverse sampling process (Song et al., 2020). Universal guidance methods allow diffusion models to be steered by user-defined guidance functions without retraining task-specific components (Bansal et al., 2023). Diffusion posterior sampling further applies gradient-based posterior guidance to general noisy inverse problems (Chung et al., 2022). Related training-free energy-guided approaches construct external energy functions from pretrained networks or attention mechanisms to control generated samples under diverse conditions (Yu et al., 2023; Park et al., 2023).

By defining an energy over luminance histograms, our method enables global, permutation-invariant control of output statistics. This allows us to manipulate HDR-relevant properties, such as dynamic range and highlight distribution, while leaving spatial structure primarily governed by the pretrained diffusion prior.

B Method Details & Proofs

B.1 Soft Histogram Consistency

Lemma 1 (Soft histogram bias)**.** *Let h be the empirical K-bin histogram of $\{Y_i\}_{i=1}^N$ and h_σ the soft histogram (3). Assume bins are uniformly spaced with width $w = 1/K$. Then*

$$\|h_\sigma - h\|_1 \;\le\; 2\sigma\sqrt{\tfrac{2}{\pi}} \cdot K \;+\; \mathcal{O}(K^{-1}) \;=\; \mathcal{O}(\sigma K + K^{-1}). \tag{9}$$

In particular, choosing $\sigma = o(1/K)$ yields vanishing bias, while the practical choice $\sigma = 0.5/K$ keeps boundary leakage controlled and preserves differentiability of h_σ in Y.

Proof. Per-bin error is bounded by the integral of a Gaussian kernel against the indicator of a width-w bin minus a width-w rectangle. Standard kernel-density bias analysis gives the leading term $2\sigma\sqrt{2/\pi}/w$ per bin and $\mathcal{O}(w)$ correction. Summing over K bins yields the stated rate. □

This motivates the practical choice $\sigma = 0.5/K$ in Section 4: it keeps the soft histogram smooth while limiting boundary leakage.

B.2 Spatial Decoupling

Corollary 1 (Equivariance of guided sampling)**.** *Let Φ_θ denote the unguided sampling map and let Φ_θ^s denote its energy-guided counterpart with energy of the form in Proposition 1. Then, for any rigid spatial transform $T \in O(2) \cap \mathrm{Stab}(\mathcal{X})$ that commutes with v_θ, $T \circ \Phi_\theta^s = \Phi_\theta^s \circ T$. Consequently, all spatial symmetries of the diffusion prior are inherited by the guided sampler.*

We empirically observe in Section 4 that LumaGuide primarily alters luminance values while largely preserving where bright and dark regions occur. The pretrained diffusion prior remains the main source of geometry and semantics, while energy guidance adjusts global feature statistics along the prescribed feature axis.

B.3 An Impossibility Result for Brightness Scaling

A natural skeptical position is that LumaGuide merely brightens images, and that a global brightness scalar would suffice. We rule this out. For a distribution P on $[0,1]$ with mean μ_P and positive standard deviation σ_P, write $\overline{P}$ for its *shape*, the law of $(Y - \mu_P)/\sigma_P$, $Y \sim P$. Two distributions share the same shape iff one is an affine reparameterization of the other.

Theorem 2 (No-go for affine luminance shaping)**.** *Let $P_0, P^* \in \mathcal{P}([0,1])$ have finite second moments and $\sigma_{P_0}, \sigma_{P^*} > 0$. For any clipped affine map $T(y) = \Pi_{[0,1]}(ay + b)$ with $a > 0$,*

$$W_1(T_\# P_0,\, P^*) \;\ge\; \sigma_{P^*}\, W_1(\overline{P_0},\, \overline{P^*}) \;-\; \rho, \tag{10}$$

with equality and $\rho = 0$ on the unclipped family, attained uniquely by the location–scale match

$$T^*(y) \;=\; \mu_{P^*} + \tfrac{\sigma_{P^*}}{\sigma_{P_0}}(y - \mu_{P_0}).$$

The clipping correction $\rho \ge 0$ vanishes whenever $T^*([0,1]) \subseteq [0,1]$. In particular, when $\overline{P_0} \ne \overline{P^*}$ — i.e. the SDR prior and the HDR target have different shapes — the right-hand side is strictly positive, and no global affine brightness adjustment can generally match the target distribution.

Proof. Affine maps preserve shape. If $Y \sim P_0$, then $aY + b$ has mean $a\mu_{P_0} + b$ and standard deviation $a\sigma_{P_0}$, so its standardized version equals $(Y - \mu_{P_0})/\sigma_{P_0}$. Hence $\overline{T_\# P_0} = \overline{P_0}$ for every unclipped T: the family cannot leave the shape orbit of P_0.

W_1 separates by shape. The 1-Wasserstein distance scales with affine reparameterizations, $W_1(b + aX, c + aY) = a\, W_1(X, Y)$ (Arjovsky et al., 2017). Combined with shape invariance, minimizing $W_1(T_\# P_0, P^*)$ over

(a, b) is a convex problem with unique solution T^*, and the minimum value equals $\sigma_{P^*} W_1(\overline{P_0}, \overline{P^*})$. This is strictly positive whenever the shapes differ, since W_1 is a metric.

Clipping costs at most ρ. The projection $\Pi_{[0,1]}$ is 1-Lipschitz, so clipping a distribution can decrease its W_1 to P^* by at most the mass it pushes outside $[0, 1]$. Defining ρ as the infimum of this excess mass over (a, b) gives (10). If $T^*([0, 1]) \subseteq [0, 1]$ no clipping is needed at T^*, hence $\rho = 0$. □ □

Theorem 2 explains the brightness-scaling baseline (Table 7) at a fundamental level: α-scaling is *provably* sub-optimal whenever the SDR-biased prior and the HDR target differ in shape, as observed empirically.

B.4 Idealized Guidance Schedule

Theorem 3 (Schedule under idealized noisy gradients)**.** *Consider the rectified flow interpolation $z_t = (1 - t)z_0 + t\epsilon$, $t \in [0, 1]$. Under (S1) bounded velocity error $\mathrm{Var}(v_\theta - v^*) \le \sigma_v^2$ uniformly in t, (S2) trajectory sensitivity scales as $(1 - t)^2$ in expected W_1-influence on the terminal sample, and (S3) endpoint variance is regularized, an influence-weighted schedule takes the form*

$$s^*(t) \;\propto\; \frac{(1-t)^2\, t^2}{\left(t^2 + (1-t)^2\right)^2}. \tag{11}$$

This schedule peaks at $t = \frac{1}{2}$.

Proof. The schedule follows by combining three idealized factors: the trajectory influence $(1 - t)^2$, the reconstruction-noise factor induced by the flow estimate, and an endpoint variance regularizer that prevents the schedule from concentrating at the boundaries. Under these assumptions, the resulting influence-to-noise weighting is proportional to

$$\frac{(1-t)^2 t^2}{\left(t^2 + (1-t)^2\right)^2}.$$

Differentiating this expression shows that its unique interior maximum occurs at $t = \frac{1}{2}$. □

Remark 1 (Why constant beats s^* in practice). Theorem 3 is derived under idealized assumptions on trajectory sensitivity and noise; in practice, mid-trajectory dynamics are dominated by the base velocity field v_θ, suppressing the effective marginal benefit of guidance there. Conversely, early-time guidance enables low-cost coarse global adjustments and late-time guidance enables high-luminance refinement. These two edge regimes are penalized by s^* and rewarded by the constant schedule, which explains the empirical hierarchy constant $\succ s^* \succ$ windowed observed in Table 6.

B.5 Temporal Stability under TLC

Proposition 3 (Bounded inter-frame luminance drift)**.** *Let $Y_\tau, Y_{\tau-1}$ be the PQ-luminance maps of adjacent LumaGuide-decoded video frames with frame-wise velocities $v_\theta^{(\tau)}$ and $v_\theta^{(\tau-1)}$. Suppose the inter-frame velocity discrepancy on the high-luminance mask satisfies $\|v_\theta^{(\tau)} - v_\theta^{(\tau-1)}\|_{\Omega_\tau} \le \kappa$. Then with the TLC term of Definition 8,*

$$\mathbb{E}\big\|Y_\tau - Y_{\tau-1}\big\|_{\Omega_\tau} \;\le\; \frac{\kappa}{\lambda_{\mathrm{TLC}}} \;+\; \mathcal{O}\!\left(\frac{\sigma_v^2}{\lambda_{\mathrm{TLC}}}\right). \tag{12}$$

The TLC-augmented energy is convex in Y_τ for fixed $Y_{\tau-1}$ and h_σ. The first-order optimality on the masked region gives $\nabla_{Y_\tau}(W_1 + \lambda_{\mathrm{TLC}}\mathcal{L}_{\mathrm{TLC}}) = 0$, i.e. $2\lambda_{\mathrm{TLC}}(Y_\tau - Y_{\tau-1}) \odot \Omega_\tau = -\nabla W_1$. Taking norms and using $\|\nabla W_1\| \le \kappa + \|\delta\|$,

$$\|Y_\tau - Y_{\tau-1}\|_{\Omega_\tau} \le \frac{\kappa + \|\delta\|}{2\lambda_{\mathrm{TLC}}}.$$

Taking expectations and using $\mathbb{E}\|\delta\|^2 \le \sigma_v^2$ yields the claim. □

B.6 Proof of Theorem 1

Let $\Phi(z) := E(\mathrm{Dec}(z - t\, v_\theta(z,t)))$ for fixed t, so that Φ is the energy as a function of the latent. By assumptions (A1)–(A2) and chain rule, Φ is L-smooth with $L = L_E\, L_{\mathrm{Dec}}^2\, (1 + t\, \|J_{v_\theta}\|)^2$. On a compact t-interval bounded away from 1, L is bounded; absorb the constant into the symbol.

Since the sampler steps from t_n to $t_{n+1} < t_n$, let $\Delta t_n := t_n - t_{n+1} > 0$. The guided update can be written as

$$\Delta z = -\Delta t_n \big(v_\theta + s \nabla \Phi\big).$$

By the descent lemma for L-smooth functions,

$$\Phi(z + \Delta z) \le \Phi(z) + \langle \nabla \Phi, \Delta z \rangle + \tfrac{L}{2} \|\Delta z\|^2.$$

Let $\eta_t := s \Delta t_n$. Substituting Δz and taking expectation over the velocity-error $\delta = v_\theta - v^*$,

$$\begin{aligned} \mathbb{E}[\Phi(z + \Delta z)] \le \Phi(z) - \eta_t\, (1 - \tfrac{L \eta_t}{2}) \mathbb{E} \|\nabla \Phi\|^2 \\ + \mathbb{E} \langle \nabla \Phi, v^* \rangle \Delta t_n + L \Delta t_n^2 \mathbb{E} \|\delta\|^2 / 2. \end{aligned}$$

The middle term is the unguided drift; under (A3) this term is bounded. We absorb $L \Delta t_n / 2 \cdot \mathbb{E}\|\delta\|^2$ into the noise term. The factor t^2 enters because, in flow-matching coordinates, $\hat{x}_0$ amplifies δ by $t/(1-t)$, contributing $t^2/(1-t)^2$ which we bound on $t \in [0, 1-\varepsilon]$ by t^2/ε^2 and absorb $1/\varepsilon^2$ into the constant C. The condition $\eta_t = s \Delta t_n \le 2/L$ guarantees the factor $1 - L\eta_t/2$ is non-negative. □

B.7 Proof of Proposition 1

(i) **Permutation equivariance.**

(ii) **Pixel-wise gradient form.**

$$\big[\nabla_x E(x)\big]_{\cdot,i} \;=\; \nabla_{x_{\cdot,i}} \varphi(x_{\cdot,i}) \cdot g\big(\varphi(x_{\cdot,i});\, h_\sigma(x), h^*\big), \tag{13}$$

(iii) **Spatial information bound.**

Proof. (i) Since $h_\sigma(x)$ is obtained by summing identical per-pixel soft assignments over all spatial locations, it is invariant to any spatial permutation $\pi \in S_N$. Hence

$$h_\sigma(\pi \cdot x) = h_\sigma(x), \qquad E(\pi \cdot x) = E(x).$$

Differentiating this identity gives

$$\nabla_x E(\pi \cdot x) = \pi \cdot \nabla_x E(x).$$

(ii) Applying the chain rule to $f(x)_i = \varphi(x_{\cdot,i})$ gives

$$\big[\nabla_x E(x)\big]_{\cdot,i} = g(\varphi(x_{\cdot,i}); h_\sigma(x), h^*)\, \nabla_{x_{\cdot,i}} \varphi(x_{\cdot,i}),$$

for some scalar function g. Thus, pixels interact through the global histogram statistics, but not through their spatial coordinates.

(iii) Since E depends on x only through the multiset

$$\{f(x)_i\}_{i=1}^N,$$

it cannot distinguish images whose feature values differ only by a spatial permutation. □

B.8 Proof of Proposition 2

Let $L \in [L_{\min}, L_{\max}]$ denote scene luminance in nits with $L_{\max}/L_{\min} \geq 10^4$ (HDR range). For a fixed bin grid in PQ space,

(i) PQ is monotone, C^∞, and $\mathrm{PQ}'(L)\,L$ is bounded above and below by positive constants on $[L_{\min}, L_{\max}]$. Hence per-pixel gradients $\partial E/\partial L$ converted to PQ space are scale-balanced across the dynamic range.

(ii) In linear space, the inverse PQ Jacobian induces polynomially mismatched scales across luminance ranges; for large L, the corresponding factor scales as $\Theta(L^{1-1/m_2})$, yielding gradient variance that grows as $\Omega(L_{\max}^{2(1-1/m_2)})$.

(iii) Consequently, the noise-to-signal ratio of the histogram-shaping gradient in PQ space is bounded by a constant independent of $L_{\max}$, while in linear space it grows polynomially with the dynamic range.

Proof. The PQ OETF is

$$\mathrm{PQ}(L) \;=\; \left(\frac{c_1 + c_2 (L/L_p)^{m_1}}{1 + c_3 (L/L_p)^{m_1}}\right)^{m_2}, \qquad L_p = 10{,}000 \text{ nits},$$

with constants $m_1 = 2610/16384$, $m_2 = 2523/4096$, $c_1 = 3424/4096$, $c_2 = 2413/128$, $c_3 = 2392/128$. Differentiating, $\mathrm{PQ}'(L) = m_2\,\mathrm{PQ}(L)^{1-1/m_2} \cdot \frac{(c_2 - c_1 c_3)\, m_1\, (L/L_p)^{m_1 - 1}/L_p}{(1 + c_3 (L/L_p)^{m_1})^2}$. For $L \gg 1$ nit, the leading scaling is $L^{m_1 - 1}$. Multiplying by L to convert to the relative scale gives $L^{m_1} \cdot \text{const}$ which, since $m_1 \approx 0.16$, is bounded above and below by constants on any range $[L_{\min}, L_{\max}]$ with $\log_{10}(L_{\max}/L_{\min}) \leq 4$. This is statement (i).

Statement (ii) follows by considering the inverse mapping from PQ perturbations to linear luminance perturbations. Since the inverse PQ Jacobian scales as $\Theta(L^{1-1/m_2})$ for large L, approximately uniform perturbations in PQ space induce polynomially mismatched luminance-scale gradients. Squaring this factor over the HDR range yields the stated variance growth.

Statement (iii) follows by combining (i) and the boundedness of the soft histogram Jacobian (whose ℓ^∞ norm is at most $1/(\sigma\sqrt{2\pi}) \cdot 1/N$ per pixel-bin pair). □ □

C Using Pretrained VAE in Perceptual Uniform Space

A key observation underlying our method is that pretrained LDR VAEs can reconstruct HDR signals with high fidelity when the inputs are expressed in a perceptually uniform space (e.g., PQ or PU21), without any retraining.

The main issue is not model capacity, but **representation mismatch**. Linear HDR is distributed very differently from LDR data: it is heavy-tailed, dominated by extreme highlights, and poorly aligned with human perceptual sensitivity. As a result, directly feeding linear HDR into an LDR-trained VAE leads to distorted latent representations and degraded reconstructions.

Perceptually uniform encodings (PQ/PU21) address this mismatch by compressing highlights and redistributing precision toward low and middle luminance regions, producing statistics that are significantly closer to LDR image distributions. Empirically, prior work (Wu et al., 2026) shows that PQ/PU21-encoded HDR can be reconstructed by pretrained VAEs with quality comparable to LDR inputs, while linear HDR fails to do so.

From a latent-space perspective, PQ encoding brings HDR signals closer to the statistics seen by LDR-trained VAEs, enabling more stable encoding and decoding without modifying the VAE.

This observation implies that HDR generation is primarily a **distribution alignment problem**. Once HDR signals are represented in a perceptually aligned space, pretrained generative models can be directly reused. Our method builds on this insight and further performs distribution-level control in PQ space, without requiring any model adaptation.

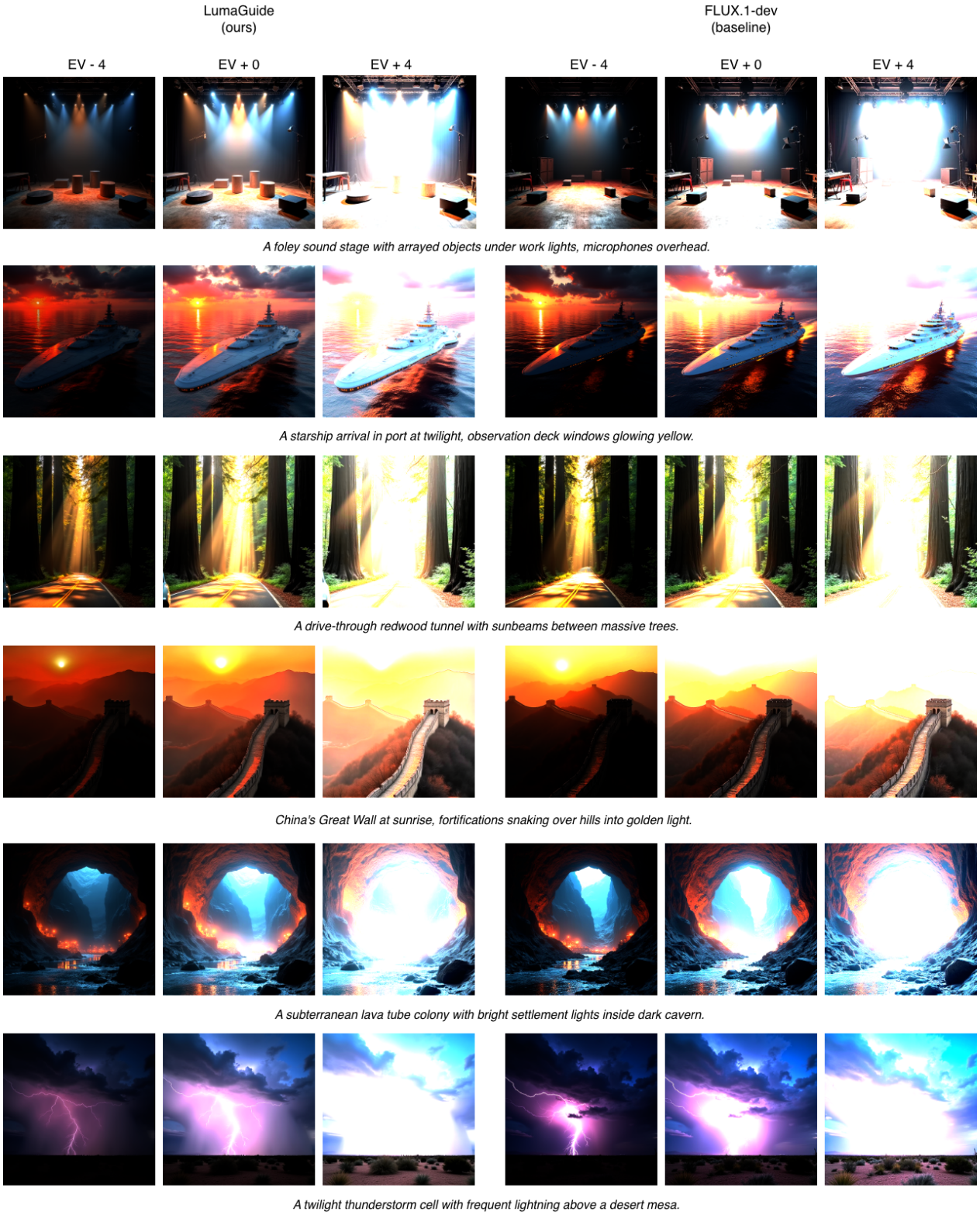


Figure 6 Additional qualitative results across diverse prompts and scenes.

D Additional Experimental Results

We provide additional qualitative results across a wider range of prompts and scenes to demonstrate the robustness of **LumaGuide**. Figure 6 shows diverse scenarios, including indoor scenes, night environments, and high-contrast outdoor settings. Across all cases, **LumaGuide** consistently reshapes luminance distributions toward HDR-like profiles while preserving semantic content. Compared to the baseline, our method better maintains characteristic HDR properties, including stronger contrast between highlights and shadows, more localized high-intensity regions (e.g., light sources, reflections, and sky highlights), and reduced overexposure in surrounding areas. In particular, bright regions remain more structured and spatially confined, rather than spreading into large saturated areas, resulting in more realistic highlight rendering.

We further evaluate the method at higher resolution using 1024×1024 generation with Flux.1-dev (Black Forest Labs, 2024). Qualitative comparisons between 512×512 and 1024×1024 are shown in Figure 7. The higher-resolution results exhibit improved texture detail and fine-grained structure, while maintaining consistent luminance distribution shaping. Notably, HDR characteristics such as highlight sharpness, contrast separation, and luminance range are preserved across resolutions, indicating that the proposed guidance operates consistently at different spatial scales.

Quantitative comparisons are summarized in Table 6, showing consistent improvements in perceptual quality metrics at higher resolution. These results suggest that **LumaGuide** generalizes across resolutions and is not restricted to a specific image scale, although the optimal guidance strength may need to be adjusted accordingly.

Table 6 Comparison between 512×512 and 1024×1024 generation using **LumaGuide**. Higher resolution improves perceptual quality, alignment, and dynamic range, at the cost of increased runtime.

Method	DR_{stops} ↑	Q-quality ↑	Q-align ↑	Time ↓
LumaGuide @ 512^2	14.99	0.568	0.814	**7.81 s**
LumaGuide @ 1024^2	**15.95**	**0.584**	**0.833**	32.3 s

E Additional Ablation Studies

E.1 Number of histogram bins.

We study the impact of histogram resolution by varying the number of bins K used to represent luminance distributions. For each value of K, we retrain the text-to-histogram regressor to predict K-dimensional PQ histograms, ensuring consistency with the binning scheme. Results are reported in Table 7.

We observe that increasing K improves fine-grained distribution matching up to $K = 64$, as reflected by lower percentile errors $p50_{\text{dist}}$ and $p99_{\text{dist}}$. However, the overall distribution distance (uW1) and perceptual quality degrade at larger K. This indicates that overly fine histogram resolution leads to diminishing returns in distribution alignment and may destabilize the optimization. We attribute this behavior to increased sensitivity to small bin-level discrepancies. As K increases, the guidance objective imposes stronger and more localized constraints, producing higher-variance gradients that can over-correct luminance values and degrade spatial coherence. Additionally, predicting high-dimensional target histograms introduces additional difficulty for the regressor, which may further contribute to performance degradation.

Overall, $K = 32$ provides a favorable trade-off between controllability and stability. It achieves strong perceptual quality and semantic alignment while maintaining competitive distribution accuracy, suggesting that intermediate histogram resolution is sufficient for capturing HDR luminance characteristics without over-constraining the optimization.

E.2 Guidance scale

We provide a more detailed analysis of the effect of guidance strength s_0 on distribution alignment and visual quality. While the main paper reports the overall trend, here we examine how different regimes of s_0 influence luminance statistics and perceptual behavior.

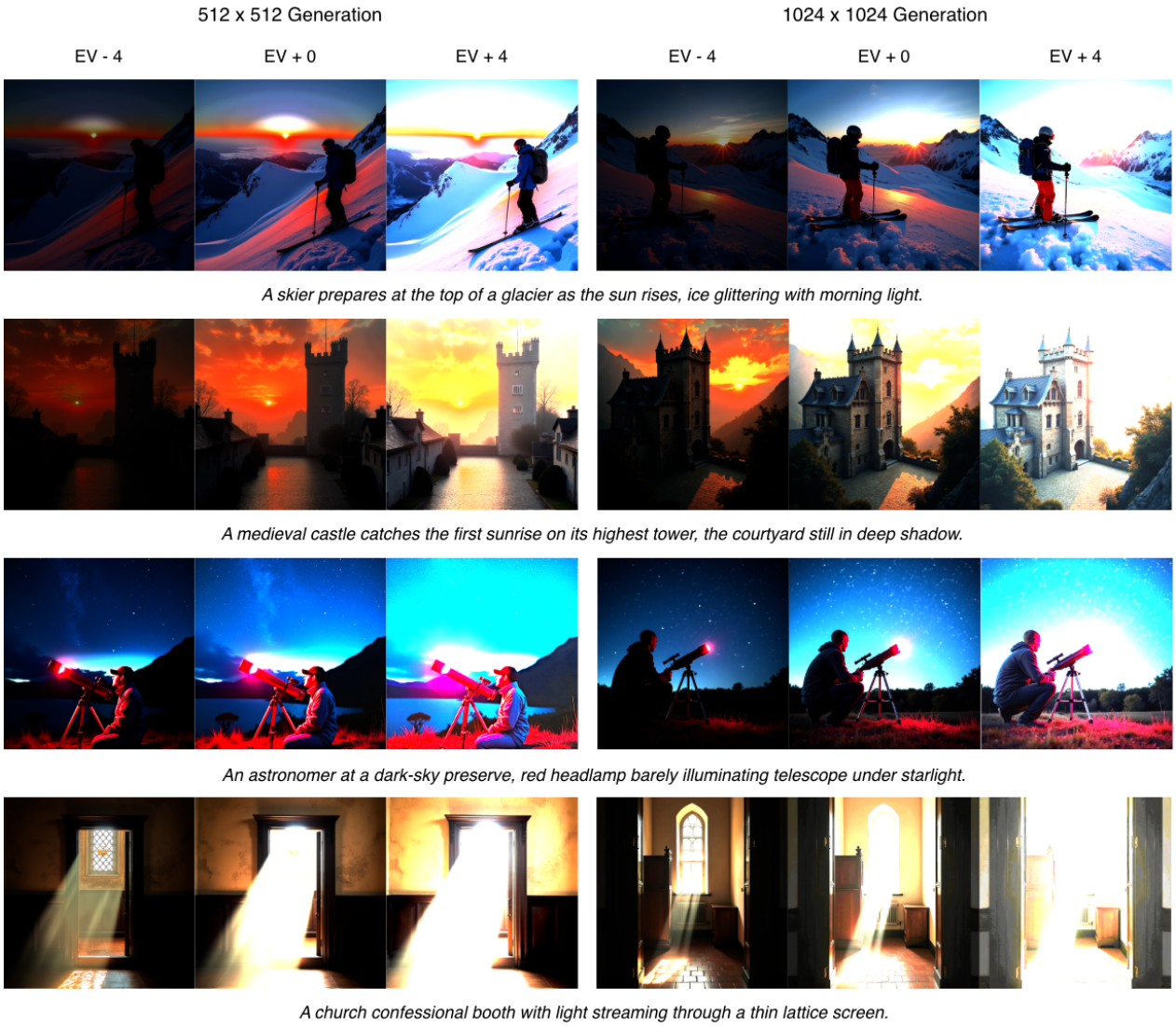


Figure 7 Resolution comparison between 512×512 and 1024×1024 generation. Higher-resolution results exhibit improved texture detail and spatial coherence while preserving consistent luminance distribution shaping.

Table 7 Effect of histogram bin count K on distribution alignment and perceptual quality. Each configuration uses a separately trained regressor to generate K-bin target distributions.

K	uW1 ↓	p50_{dist} ↓	p99_{dist} ↓	DR_{stops} ↑	Q-quality ↑	Q-alignment ↑
16	0.696	0.056	0.113	**15.48**	0.554	**0.823**
32	**0.576**	0.024	0.053	14.99	**0.568**	0.814
64	0.908	**0.015**	**0.033**	14.54	0.533	0.766
128	2.317	0.017	0.037	14.52	0.490	0.684
256	4.734	0.019	0.039	14.26	0.371	0.470

Table 8 Ablation of guidance schedules under integral normalization.

Schedule	Form	uW1 ↓	p50_{dist} ↓	p99_{dist} ↓	DR ↑	Q-quality ↑	Q-align ↑
constant	$s(t) = s_0$	**0.576**	**0.024**	**0.053**	**14.99**	**0.568**	**0.814**
snr_weighted	bell-shaped	0.979	0.037	0.079	14.94	0.538	0.810
late_only	windowed	1.484	0.049	0.105	14.96	0.514	0.787

As shown in Table 1, increasing s_0 generally reduces distribution discrepancy up to an intermediate range, as measured by the Wasserstein-1 (W_1) distance. This confirms that the guidance signal effectively steers the generated samples toward the target luminance distribution. Improvements are observed across both global (uW1) and percentile-based metrics ($p50_{\text{dist}}$, $p99_{\text{dist}}$), indicating that the alignment affects the full distribution rather than a specific range.

At moderate values (e.g., $s_0 \in [500, 2000]$), the constraint is strong enough to redistribute luminance mass from extreme highlights toward mid-tones, improving perceptual balance and reducing overexposure. At higher values of s_0, however, the constraint becomes overly dominant relative to the diffusion prior. As a result, the optimization becomes increasingly sensitive to small distribution discrepancies, leading to over-correction and degradation of spatial coherence.

Figure 9 further illustrates this trade-off. At low s_0, outputs remain close to the baseline distribution and exhibit limited HDR characteristics. As s_0 increases, the outputs progressively align with the target distribution, improving highlight structure and shadow detail. Beyond a certain point, further increases in s_0 degrade perceptual quality due to over-constrained optimization. Based on this trade-off, we select $s_0 = 2000$ as the default operating point, since it achieves strong alignment while preserving perceptual quality and avoiding the over-constrained behavior observed at larger scales.

E.3 Guidance schedule

We analyze the role of the guidance schedule $s(t)$ in distribution shaping. Consider the rectified flow interpolation

$$z_t = (1-t)z_0 + t\epsilon, \qquad t \in [0,1],$$

with signal-to-noise ratio $\text{SNR}(t) = \frac{(1-t)^2}{t^2}$. The guided update is

$$v_{\text{guided}}(z_t, t) = v_\theta(z_t, t) + s(t)\,\nabla_{z_t} E.$$

We first consider an idealized schedule motivated by expected downstream energy reduction. Two factors govern the effectiveness of guidance at time t:

(i) Gradient reliability. Let $\hat{z}_0 = z_t - t\,v_\theta(z_t, t)$. Writing $v_\theta = v + \delta$ with prediction error δ, we have

$$\hat{z}_0 = z_0 - t\,\delta.$$

Assuming $\text{Var}(\delta) \propto 1$, the variance of the reconstruction scales as $\text{Var}(\hat{z}_0) \propto t^2$, and hence

$$\text{Var}(\nabla E) \propto t^2.$$

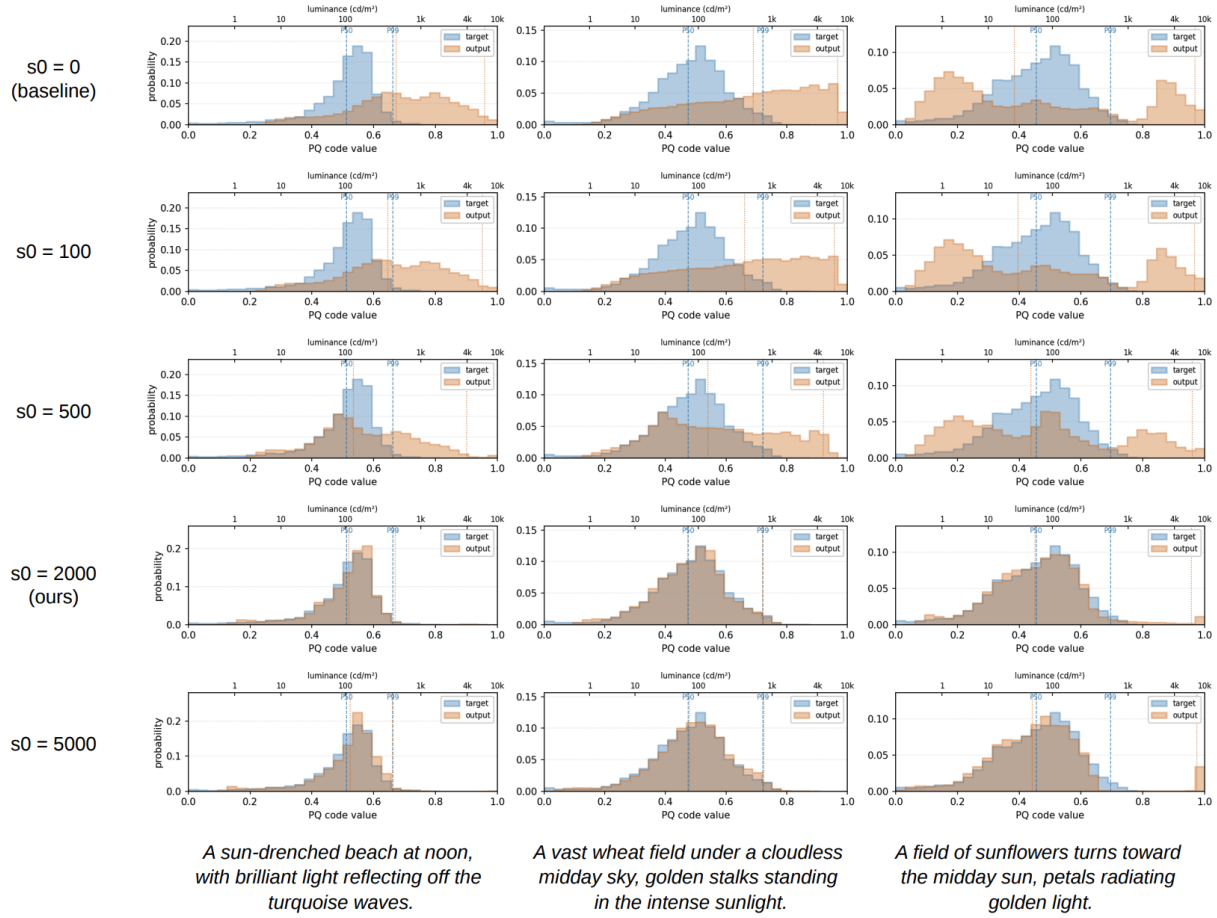


Figure 8 Luminance histogram alignment in PQ space. **LumaGuide** reshapes the output distribution toward the target HDR profile.

(ii) Trajectory sensitivity. A perturbation at time t propagates to the final sample with magnitude proportional to the remaining trajectory length. In rectified flow, this scales as $(1-t)$, giving an effective sensitivity $(1-t)^2$.

Combining these two factors suggests the scaling

$$s^*(t) \;\propto\; \frac{(1-t)^2}{t^2}.$$

Introducing a variance floor and normalizing leads to the bounded form

$$s^*(t) \;\propto\; \frac{(1-t)^2 t^2}{\left(t^2+(1-t)^2\right)^2},$$

which peaks at $t = 0.5$.

We compare this SNR-weighted schedule with a constant schedule and a time-windowed variant. All schedules are normalized to have equal $\int s(t)\,dt$. Results are shown in Table 8.

Empirically, the constant schedule outperforms all alternatives across both distributional and perceptual metrics. The SNR-weighted schedule yields nearly 2× higher uW1, and the windowed schedule performs worst overall.

The discrepancy arises because the assumptions above do not hold uniformly across the trajectory. In practice, mid-range timesteps are dominated by the base flow dynamics, reducing the effective impact of

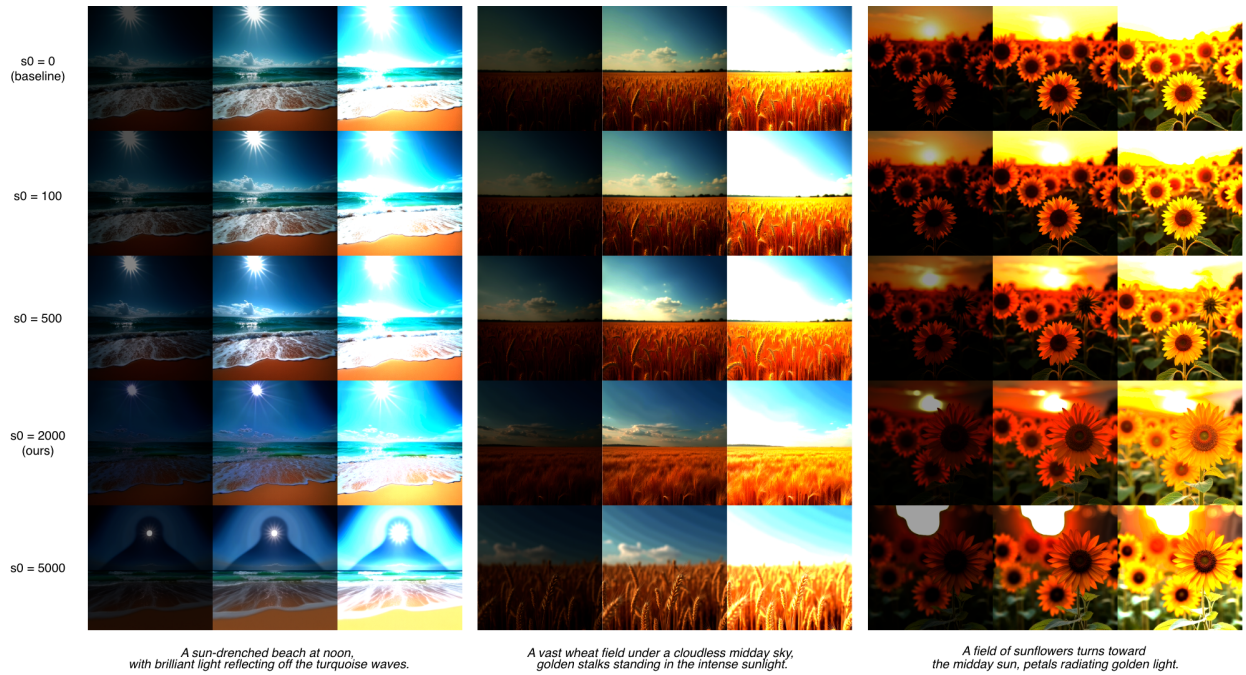


Figure 9 Effect of guidance strength s_0 on luminance distribution and visual appearance. Each row corresponds to a different guidance scale, and columns show representative scenes under increasing exposure levels.

Table 9 Comparison of PQ and PU21 encoding for distribution shaping at different resolutions.

Encoding	Res	DR (stops) ↑	Q-quality ↑	Q-alignment ↑
PQ	512	14.99	**0.568**	**0.814**
PU21	512	**15.95**	0.556	0.813
PQ	1024	16.02	0.584	0.833
PU21	1024	**16.40**	**0.588**	**0.834**

guidance. Conversely, early timesteps allow low-cost global adjustments, while late timesteps remain sensitive for high-luminance refinement. As a result, concentrating guidance near $t = 0.5$ is suboptimal.

A constant schedule distributes guidance uniformly across the trajectory, enabling both global and local adjustments, and yields the best overall performance.

E.4 Effect of HDR encoding (PQ vs. PU21)

We compare two perceptually motivated HDR encoding schemes, PQ and PU21, for distribution shaping. Results are summarized in Table 9.

PU21 consistently achieves higher dynamic range across resolutions, indicating stronger expansion in extreme luminance regions. It also provides slightly better perceptual quality and alignment at higher resolution. However, PQ remains competitive across all settings and achieves stronger perceptual quality at lower resolution. More importantly, PQ is a widely adopted and standardized HDR encoding, making it more suitable for general-purpose use.

Overall, both encodings are effective for distribution shaping. We adopt PQ as the default due to its robustness and broader compatibility, while noting that PU21 may provide advantages in extreme dynamic range scenarios.

Table 10 Comparison of backbone characteristics used for cross-backbone evaluation. The three models span different architectures, scheduler families, and VAE designs, enabling controlled analysis of distribution shaping behavior.

Property	Flux.1-dev	SDXL	SD3-medium
Backbone	MMDiT	UNet	MMDiT
Scheduler	Rectified flow	ϵ-prediction	Rectified flow
Parameters	12B	3.5B	2B
Text encoder	T5 + CLIP	CLIP (dual)	T5 + CLIP
VAE channels	16	4	16

Table 11 Cross-backbone comparison of **LumaGuide** at the selected operating points for each model. Metrics report perceptual quality, alignment, and dynamic range.

Backbone	Q-quality ↑	Q-align ↑	DR (stops) ↑
Flux (ours)	**0.568**	**0.814**	14.99
SD3	0.512	0.795	15.42
SDXL	0.431	0.655	**15.90**

F Additional Baselines

A key question for distribution shaping methods is whether the observed behavior is intrinsic to the method or tied to a specific backbone. To establish that **LumaGuide** is architecture-agnostic, we evaluate it across three representative diffusion backbones: Flux.1-dev, SD3-medium, and SDXL base 1.0. Their key characteristics are summarized in Table 10. These models differ along multiple axes, including architecture (MMDiT vs. UNet), scheduler family (rectified flow vs. ϵ-prediction), VAE design (16-channel native vs. 4-channel sRGB-trained), and model scale. This selection allows us to assess whether W_1-guided distribution shaping remains effective across substantially different model designs.

The three backbones are chosen to span the design space of modern diffusion models. Flux.1-dev serves as the strongest open-source baseline and our default backbone. SD3-medium shares the same rectified-flow formulation but differs in scale and training distribution, allowing us to isolate whether the method depends on model size or data. SDXL, in contrast, represents a previous-generation UNet-based model with ϵ-prediction and an sRGB-trained VAE, providing a maximal shift in both architecture and latent representation. Together, these models provide a stress test of how scheduler family, architecture, and VAE design affect distribution shaping behavior.

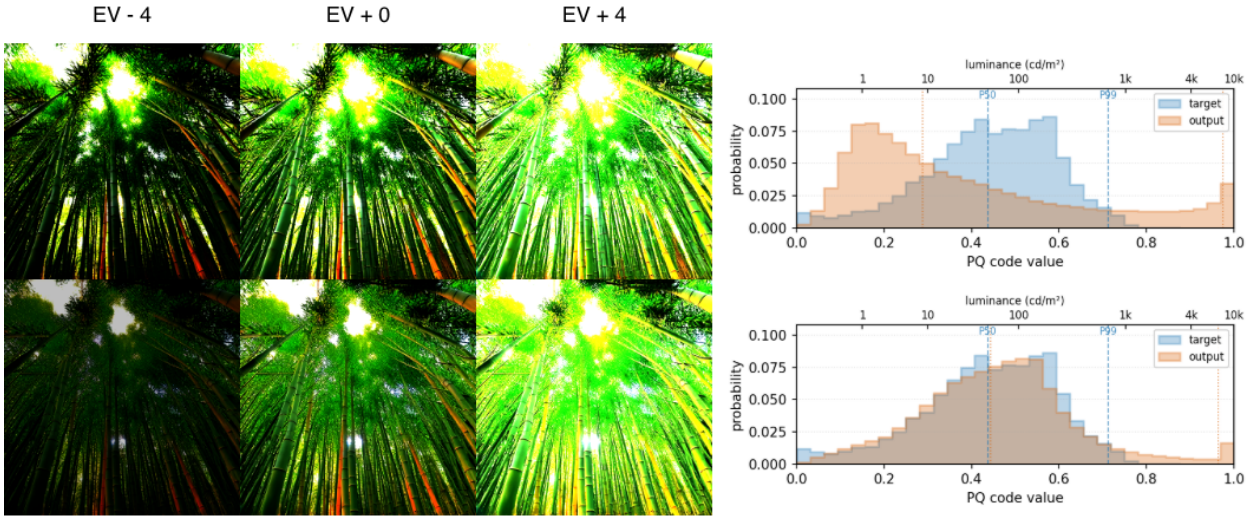
F.1 Results on SD3.

We first evaluate **LumaGuide** on SD3-medium. Quantitative results are summarized in Table 11, and qualitative comparisons are shown in Figure 10. Similar to Flux.1, increasing the guidance scale consistently improves distribution alignment on SD3, as shown in Table 12. The method consistently reduces distribution error (uW1 and percentile distances). Qualitatively, the generated images exhibit improved highlight structure and more consistent HDR-specified luminance distributions, as evidenced by the closer alignment between output and target histograms.

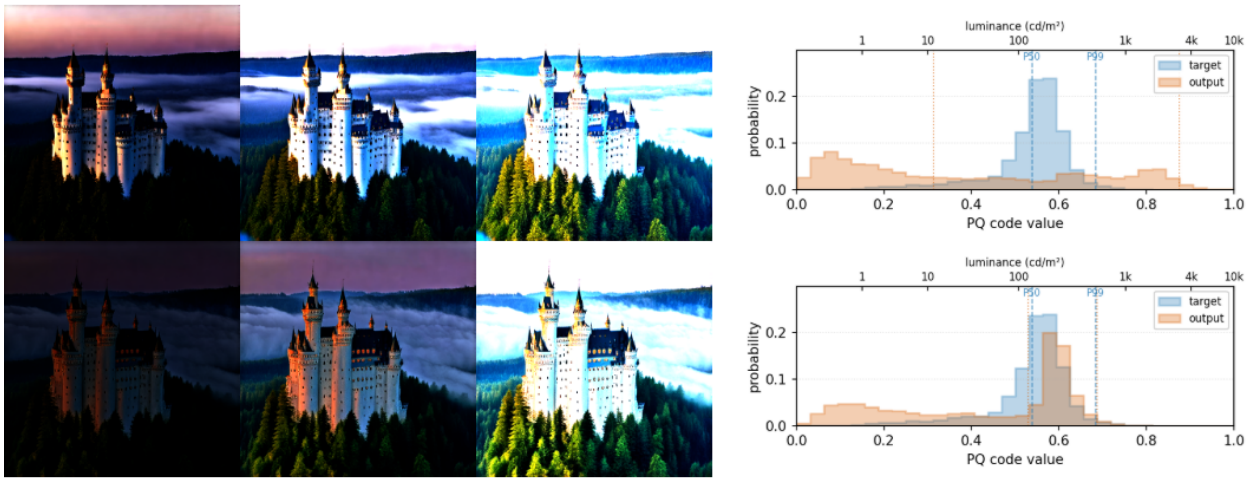
Consistent with observations on Flux.1, overly strong guidance leads to visible artifacts, such as banding and unnatural textures. We therefore select $s_0 = 1000$ as a balanced operating point for SD3, since it achieves strong distribution alignment while avoiding the artifacts observed at larger scales.

F.2 Results on SDXL.

We next evaluate **LumaGuide** on SDXL, which differs substantially from Flux and SD3 in both architecture and scheduler. Results are reported in Table 11, with qualitative examples in Figure 11. Despite these differences, the same qualitative behavior is observed: increasing guidance strength improves distribution



A bamboo forest at midday, vertical sunbeams cutting downward between green stalks.



Germany's Neuschwanstein castle at dawn, towers rising above forest mist.

Figure 10 Qualitative results on SD3 and luminance distribution matching at $s_0 = 1000$.

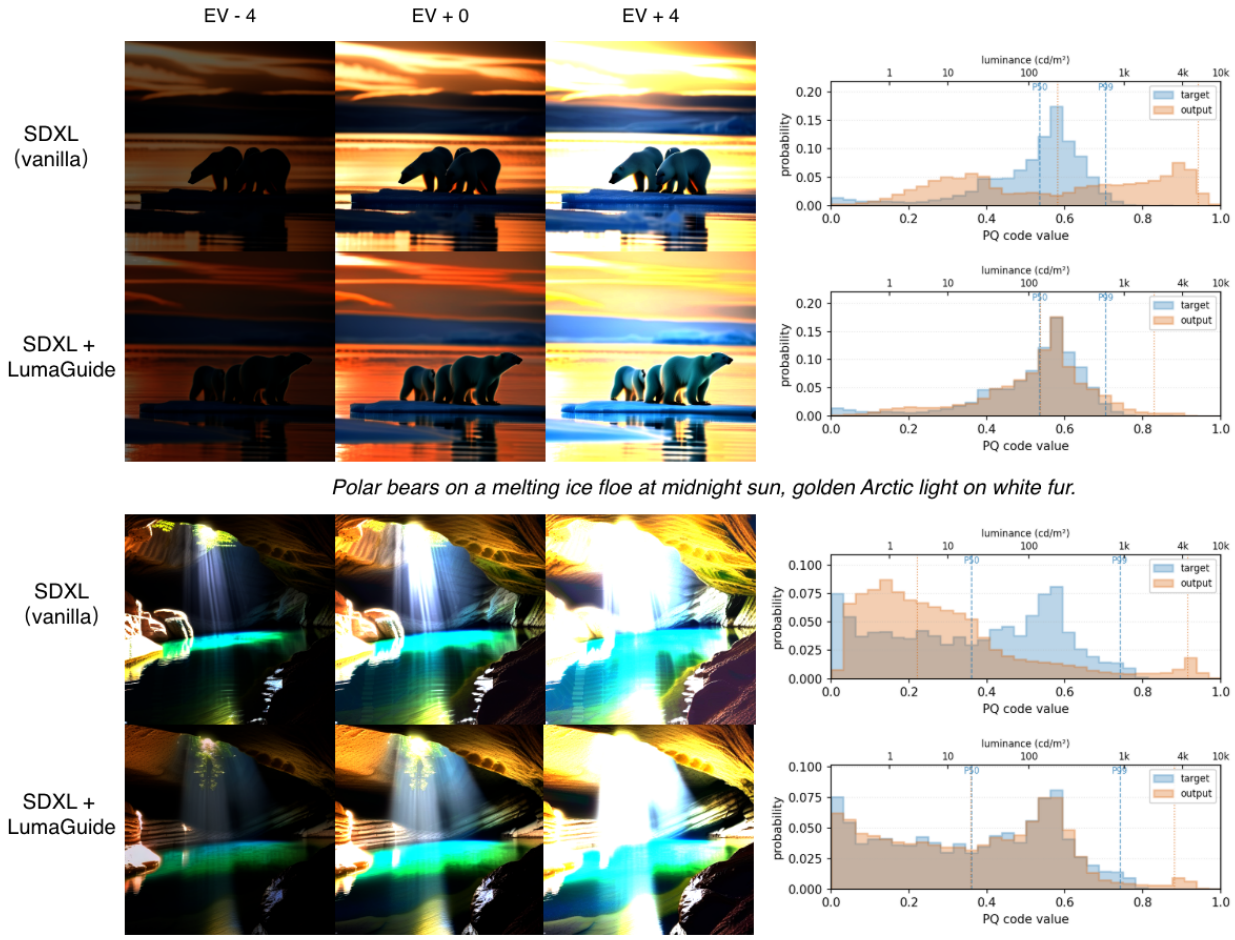


Figure 11 Qualitative results on SDXL and luminance distribution matching at $s_0 = 50$.

Table 12 Effect of guidance scale s_0 on distribution alignment for SD3-medium.

s_0	uW1 ↓	p50_{dist} ↓	p99_{dist} ↓	nits_{50}	nits_{99}
0	4.358	0.130	0.251	83.03	6708.69
100	3.755	0.110	0.246	62.83	6657.08
500	2.121	0.066	0.199	36.80	5298.82
1000	1.145	0.038	0.149	37.31	4292.97
2000	0.762	0.028	0.120	40.84	3944.31

Table 13 Effect of guidance scale s_0 on distribution alignment for SDXL.

s_0	uW1 ↓	p50_{dist} ↓	p99_{dist} ↓	nits_{50}	nits_{99}
0	4.140	0.124	0.236	40.66	6349.08
10	3.291	0.092	0.229	28.30	6267.76
50	1.470	0.032	0.200	34.92	5817.46
100	0.907	0.022	0.168	38.52	5098.24
200	0.689	0.017	0.187	38.37	6141.69

alignment (Table 13) but may introduce mild artifacts. We select $s_0 = 50$ as the operating point for SDXL because larger scales improve global distribution metrics but introduce more visible artifacts and do not fully resolve the highlight-tail gap.

However, two notable differences emerge. First, the selected guidance scale is significantly smaller than in rectified-flow models, reflecting differences in the noise schedule and step size. Second, SDXL exhibits a persistent gap in highlight alignment, with higher $p99_{\text{dist}}$ compared to Flux and SD3.

G Details of the Text-to-Histogram Regressor

Training data. We use a large-scale HDR user-generated content (UGC) dataset, Beyond8bits (Saini et al., 2026a, 2025a, 2026b). From each video, we uniformly sample four frames, resulting in about 24,000 unique HDR images. All frames are stored in BT.2020 color space with ST.2084 (PQ) encoding, represented as RGB values in $[0, 1]$.

For each frame, we generate a free-form natural language caption using Qwen2.5-VL-7B-Instruct (Team, 2024), prompting it to describe scene content, lighting conditions, and dominant materials. The resulting (caption, frame) pairs provide supervision for learning a mapping from text to luminance distributions.

Target representation. For each frame, we compute a 32-bin soft histogram of PQ luminance over the $[0, 1]$ range. Soft binning is implemented using a Gaussian kernel with standard deviation $\sigma = 0.5/K$ (half a bin width, where $K = 32$). This yields a normalized probability vector representing the luminance distribution.

Model architecture. The regressor maps a text caption to a luminance histogram. Captions are first encoded using a frozen CLIP ViT-L/14 text encoder, producing a 768-dimensional embedding. This embedding is then passed through a three-layer MLP with dimensions $768 \to 256 \to 256 \to 32$, using GELU activations. A softmax layer is applied at the output to produce a normalized distribution. The model contains approximately 270K trainable parameters, with the text encoder kept frozen.

Training protocol. We train the regressor by minimizing the KL divergence between the target and predicted histograms, $\text{KL}(\text{target} \,\|\, \text{prediction})$, averaged over the batch. Optimization is performed using AdamW with learning rate 10^{-3} and weight decay 10^{-4}, for 50 epochs with batch size 256.

To ensure proper generalization, we split the dataset at the video level (i.e., frames from the same video are assigned to the same partition), using a 90% / 5% / 5% train/validation/test split. The final model is selected based on validation KL divergence.

Results. On the held-out test set (video-disjoint), the regressor achieves a KL divergence of 0.346, close to the best validation performance (0.338), indicating good generalization to unseen content. These results suggest that text descriptions provide sufficient signal to predict coarse luminance distributions, enabling flexible text-driven specification of target histograms.

H Subjective Study

Setup We conducted a four-way ranking subjective study comparing our method against three state-of-the-art HDR generation baselines: BracketDiffusion (Bemana et al., 2025), LEDiff (Wang et al., 2025), and X2HDR (Wu et al., 2026). The study was conducted on consumer-level HDR displays. Stimuli were presented through a Chrome web browser in full-screen mode.

Stimuli We used the generated output images spanning diverse HDR-relevant lighting scenarios, including specular highlights, low-light scenes with point sources, high-contrast outdoor environments, dusk/twilight conditions, and indoor mixed lighting. For each prompt, we generated outputs from all four methods using a fixed random seed.

To enable fair perceptual comparison while controlling for absolute luminance differences across methods, all outputs were normalized following the convention of X2HDR: each image was scaled such that its 99.5th-percentile luminance mapped to 4000 cd/m^2. Images were encoded as 10-bit AVIF and verified for correct HDR metadata before the study.

Protocol We adopted a four-alternative full-ranking protocol. On each trial, four images (one per method) were presented simultaneously in a randomized 2×2 grid, and observers ranked them from best (1) to worst (4) based on overall HDR quality, considering realism of bright highlights, detail preservation in shadows, naturalness of color and contrast, and absence of artifacts (e.g., banding, blown highlights, color shifts). Observers were explicitly instructed to disregard text-prompt alignment and focus solely on HDR image quality.

Each observer completed 3 practice trials (data discarded) followed by 40 main trials, with image positions randomized per trial to control for position bias. A 600 ms mid-gray inter-trial blank was inserted between trials to reset visual adaptation, and a 30-second mid-session rest was offered after the 20th trial. Sessions lasted approximately 15 minutes depending on observer deliberation time.

Observers We recruited 15 observers with normal or corrected-to-normal vision and normal color vision. Expertise levels ranged from naive viewers to HDR/imaging experts.

Analysis For each trial, the 4-way ranking yields $\binom{4}{2} = 6$ pairwise outcomes. We converted the 600 rankings into $600 \times 6 = 3600$ pairwise outcomes and aggregated them into a 4×4 pairwise preference matrix. Method scores were estimated via maximum-likelihood Bradley–Terry and converted to Just Objectionable Difference (JOD) units, where 1 JOD corresponds to a 75% pairwise preference threshold. 95% confidence intervals were computed via 2000-iteration bootstrap resampling over trials.

Results. Table 14 summarizes the overall ranking results. **LumaGuide** achieves the best subjective performance, with the lowest mean rank, highest win rate, highest Top-2 rate, lowest worst-rank rate, and highest JOD score among all methods. Compared with X2HDR, the strongest baseline, **LumaGuide** obtains a higher win rate and a lower worst-rank rate, indicating fewer severe failure cases and more consistently preferred HDR outputs across diverse content.

I Additional Video Results

We evaluate the generality of LumaGuide on video generation using CogVideoX-5B, a 5B-parameter open-source text-to-video latent diffusion model. The evaluation uses 30 prompts covering different types of scenes.

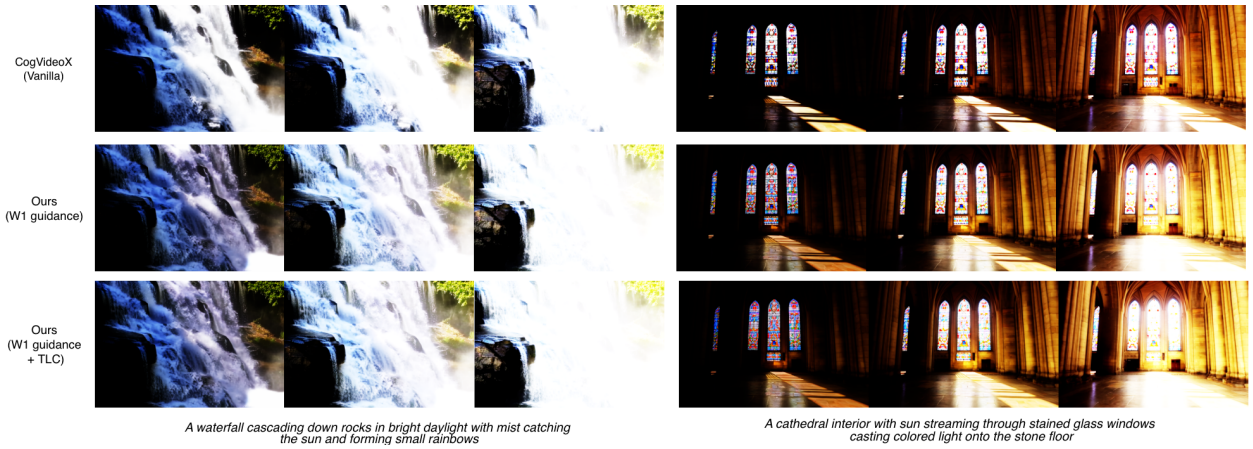


Figure 12 Qualitative video examples comparing vanilla CogVideoX, W_1-guided LumaGuide, and LumaGuide with TLC. LumaGuide improves highlight control and luminance consistency while preserving scene content.

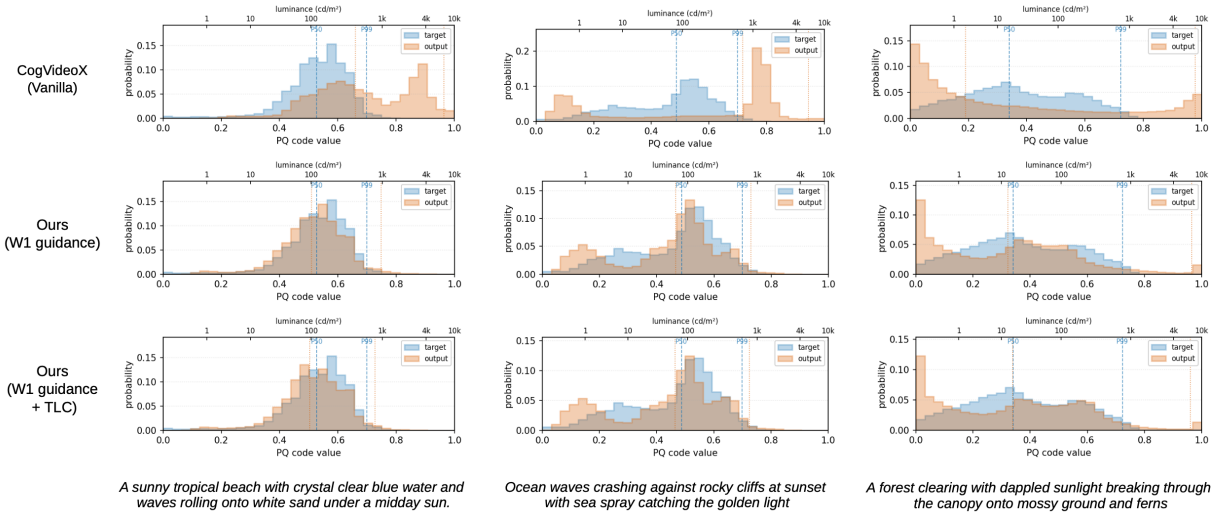


Figure 13 Distribution alignment of generated video under three representative prompts. Top row: Vanilla CogVideoX (Yang et al., 2024) produces over-saturated PQ histograms that are poorly aligned with the target distribution (blue), with excessive mass concentrated in high-luminance regions. Middle row: Our W_1 guidance shifts the output distributions (orange) toward the target, improving alignment across both mid-tones and high-intensity regions while preserving semantic content. Bottom row: Adding temporal luminance coherence (TLC) maintains distribution alignment.

Table 14 Overall subjective ranking results. Lower mean rank and worst-rank rate are better; higher win rate, Top-2 rate, and JOD are better.

Method	Mean rank ↓	Win rate ↑	Top-2 rate ↑	Worst rate ↓	JOD [95% CI]
LumaGuide	**1.80**	**50.2%**	**77.7%**	**7.7%**	**+0.75 [+0.65, +0.85]**
X2HDR	2.09	32.0%	71.7%	12.8%	+0.43 [+0.34, +0.52]
BracketDiffusion	2.79	12.5%	32.5%	24.5%	−0.30 [−0.39, −0.21]
LEDiff	3.32	5.3%	17.7%	55.2%	−0.88 [−0.99, −0.79]

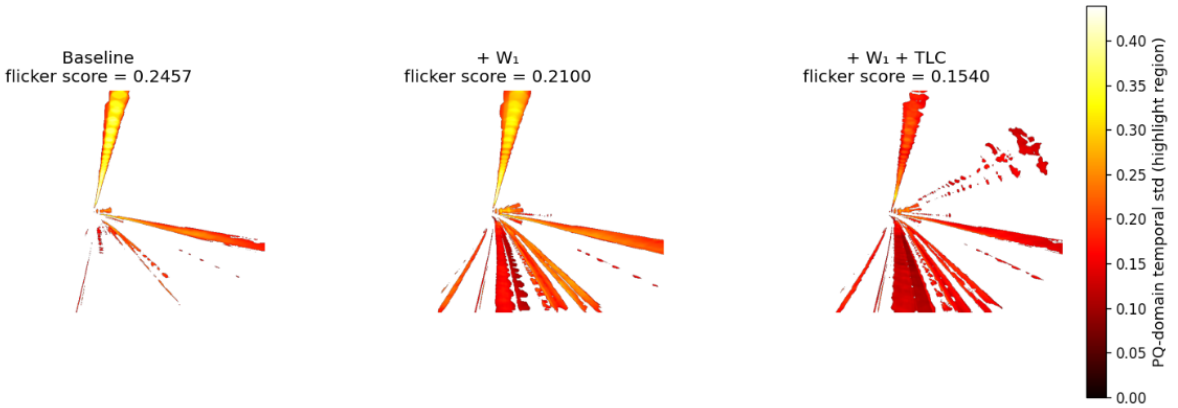


Figure 14 Per-pixel temporal standard deviation in highlight regions (PQ > 0.6) for the prompt *A subway tunnel with a train approaching and bright headlights brightening the dark concrete walls.* The baseline exhibits substantial flicker (yellow regions, score 0.246). W_1 guidance alone marginally reduces flicker (0.210). Adding TLC further suppresses flicker, as the temporal-variance term explicitly penalizes per-pixel variation in high-luminance regions.

For each prompt, we generate a 33-frame video at 480×720 resolution (8 FPS, ∼4 seconds) with 50 diffusion steps. We compare three settings: (a) the vanilla CogVideoX baseline, (b) adding PQ-domain W_1 guidance ($s_0 = 3000$, constant schedule, perceptual-log formulation), and (c) adding temporal luminance coherence (TLC) on top of W_1 ($\lambda_f = \lambda_c = 10$). Target luminance histograms are predicted using our text-to-histogram regressor. All experiments are conducted on a single NVIDIA GH200 GPU.

Since there is no widely adopted reference-free metric for HDR video generation, we rely on two simple, ground-truth-free measures. First, uW_1 measures the Wasserstein-1 distance between the generated PQ luminance histogram (aggregated across frames) and the predicted target distribution. Second, we compute flicker variance as the mean temporal variance of per-pixel luminance in highlight regions (PQ > 0.6), which captures instability in bright areas.

Figure 13 shows that the baseline produces poorly aligned luminance distributions, with a large portion of mass pushed toward high PQ values (near 1.0), which corresponds to over-saturated highlights. Adding W_1 guidance consistently shifts the distributions toward the target across all prompts, improving both median (P50) and high-percentile (P99) alignment, in line with the drop in uW_1 in Table 15.

This effect is also visible in Figure 12. The baseline tends to over-expose bright regions and produces noticeable frame-to-frame fluctuations. With W_1 guidance, highlight structure becomes more controlled and luminance is redistributed more evenly, while the overall scene content is preserved.

Figure 14 highlights the temporal behavior. The baseline shows large high-variance regions in bright areas, indicating strong flicker. W_1 alone reduces this only slightly. When TLC is added, these high-variance regions are largely suppressed, leading to visibly more stable highlights across frames. This is consistent with the reduction in flicker variance from 0.025 to 0.016.

Importantly, adding TLC does not harm distribution alignment, as uW_1 remains essentially unchanged. Overall, W_1 mainly improves the spatial luminance distribution, while TLC addresses temporal instability, and the two components work well together.

Table 15 Quantitative results on CogVideoX-5B.

Method	uW$_1$ ↓	P50 gap ↓	P99 gap ↓	Flicker var ↓
Vanilla CogVideoX	6.85 ± 2.83	0.250	0.245	0.033
+ W_1 guidance	4.03 ± 3.08	0.159	0.155	0.025
+ W_1 + TLC	**3.98** ± 3.12	0.160	**0.154**	**0.016**

Table 16 Pairwise improvements across prompts.

Comparison	Mean Δ flicker var	Median Δ
baseline → + W_1	0.008	0.007
+ W_1 → + W_1 + TLC	0.009	0.006
baseline → + W_1 + TLC	0.018	0.015

J Analysis on Evaluation Metrics

We measure dynamic range following (Wu et al., 2026). For each generated image, we first compute BT.2020 weighted luminance in linear HDR space:

$$Y(x, y) = 0.2627R + 0.6780G + 0.0593B, \tag{14}$$

where Y is measured in cd/m^2. To suppress isolated decoder noise that may artificially inflate luminance extremes, we apply Gaussian smoothing with $\sigma = 3$ pixels before computing statistics.

We then estimate the effective dynamic range using the 0.5th and 99.5th luminance percentiles, which robustly capture the visible luminance span while excluding numerical outliers. Following common HDR evaluation practice, the lower percentile is clamped to 0.05 cd/m^2 to avoid unstable ratios near black.The final metric is defined as:

$$\mathrm{DR}_{\mathrm{stops}} = \log_2 \left(\frac{\tilde{Y}_{99.5}}{\max(\tilde{Y}_{0.5}, 0.05)} \right), \tag{15}$$

where $\tilde{Y}$ denotes the Gaussian-smoothed luminance.

We also adopt the Q-Eval quality and alignment scores following prior HDR generation work. However, we observe a systematic discrepancy between Q-Eval quality scores (on HDR) and perceptual assessment in our setting.

In our framework, generated images are represented in PQ space, and evaluation is performed directly on PQ-encoded signals following prior HDR generation work. As a result, the input distribution to Q-Eval differs from its expected input regime, which is closer to SDR-like image statistics. This introduces a mismatch between the evaluation metric and the target signal domain. We observe that increasing the guidance scale generally decreases Q-Eval quality scores. At large guidance values, this behavior is expected, as overly strong guidance introduces visible artifacts such as banding and unnatural textures. However, even within the moderate guidance regime, where visual inspection indicates improved HDR characteristics and better alignment with the target luminance distribution, Q-Eval quality scores still decrease.

This suggests that Q-Eval quality is sensitive to deviations from its learned SDR prior, rather than to the correctness of HDR luminance distributions. In particular, the redistribution of luminance toward HDR distribution, which is essential for HDR generation, may be penalized as a distribution shift from SDR statistics.

In contrast, the Q-Eval alignment score remains relatively stable across guidance scales. This is consistent with its design, as it primarily measures semantic consistency between the generated image and the text prompt, and is less sensitive to luminance distribution.

These observations highlight a limitation of current evaluation metrics when applied to HDR generation. While Q-Eval provides a useful proxy for perceptual quality, it does not fully capture the correctness of HDR luminance structure. Therefore, we complement it with distribution-based metrics and qualitative analysis

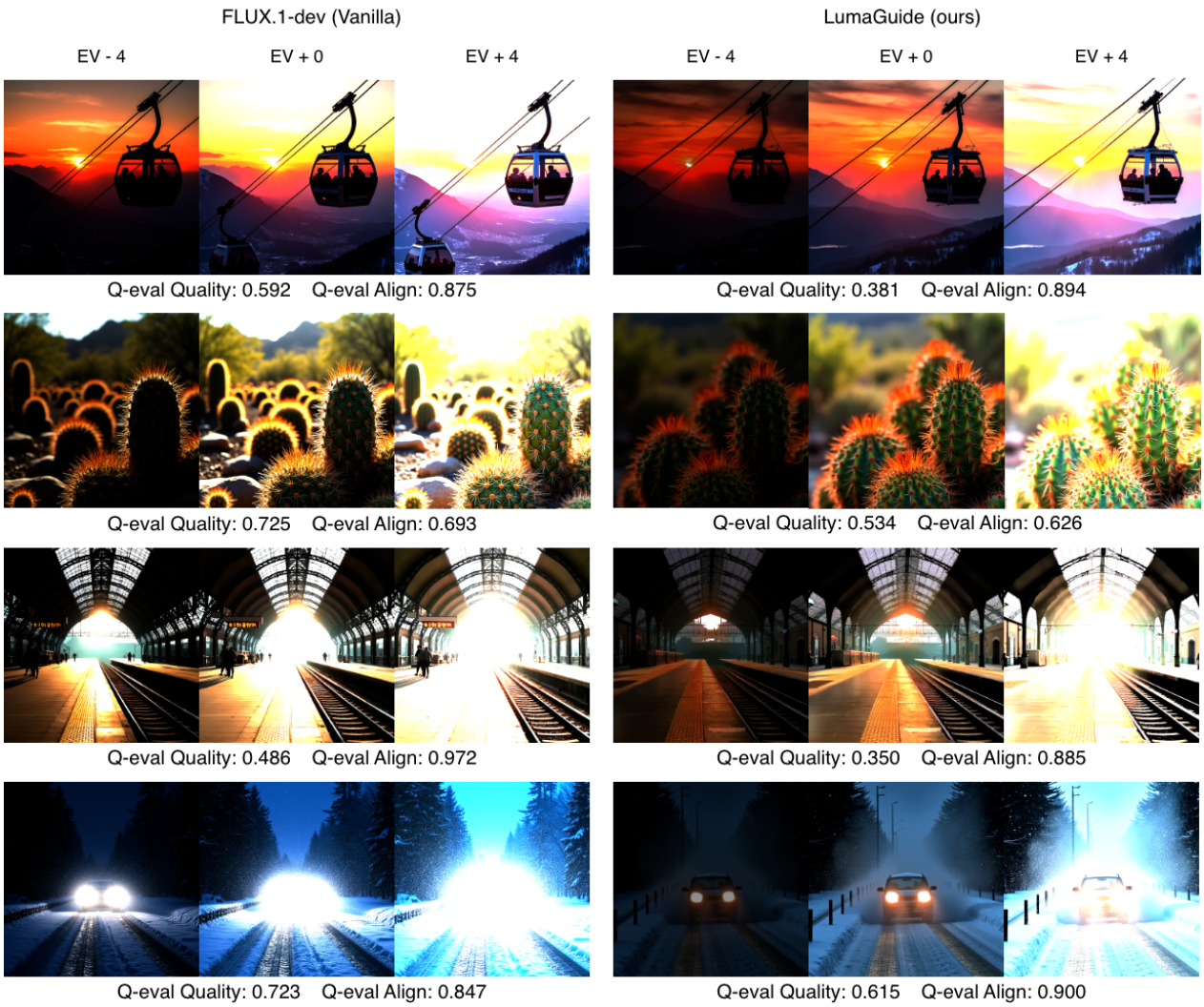


Figure 15 Illustration of the mismatch between Q-Eval quality scores and perceptual HDR quality. Images with stronger HDR luminance structure may receive lower Q-Eval quality scores despite improved distributional alignment.

in our evaluation. As shown in Figure 15, this discrepancy is clearly visible: images with more realistic HDR behavior receive lower Q-Eval quality scores.

K Failure Cases and Limitations

We analyze two representative failure modes of **LumaGuide** that arise from the interaction between distribution-level guidance and the underlying generative model.

Saturation-Induced Unguidable Regions. In scenes dominated by strong light sources (e.g., sky, sun, specular highlights), we observe that some pixels remain saturated at the maximum PQ value and cannot be reduced by guidance shown in Figure 16. Increasing the guidance strength s_0 often darkens surrounding regions while leaving saturated pixels unchanged, resulting in contrast exaggeration and halo artifacts. This behavior is caused by gradient vanishing at the decoder output: once pixels are clamped to the upper bound, their gradients become zero and no longer propagate back to the latent variables. As a result, the Wasserstein objective only affects neighboring unsaturated pixels, creating a luminance imbalance. This limitation prevents full alignment with the target distribution and leads to residual errors in high-percentile statistics (e.g., $p99$). Addressing this issue may require smoother decoding functions or HDR-aware architectures that preserve

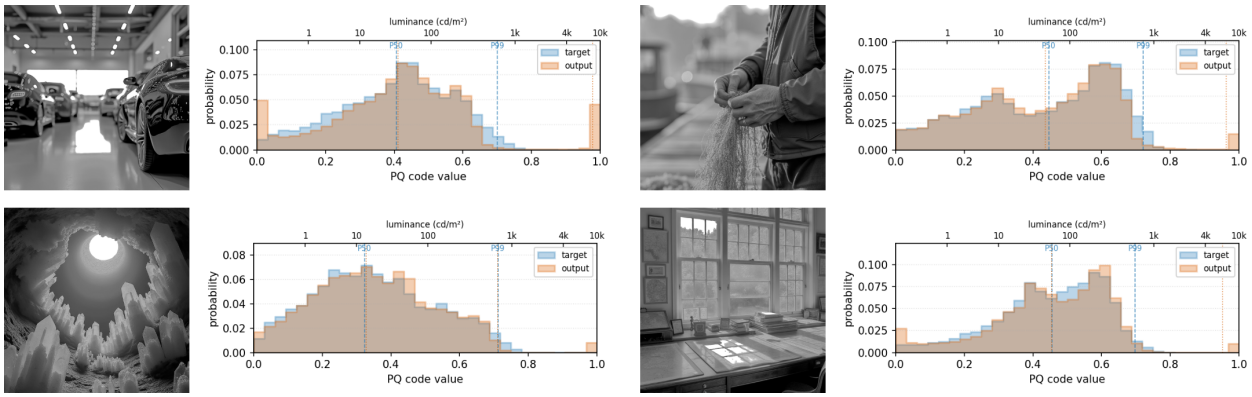


Figure 16 Failure case involving saturation-induced unguidable regions.

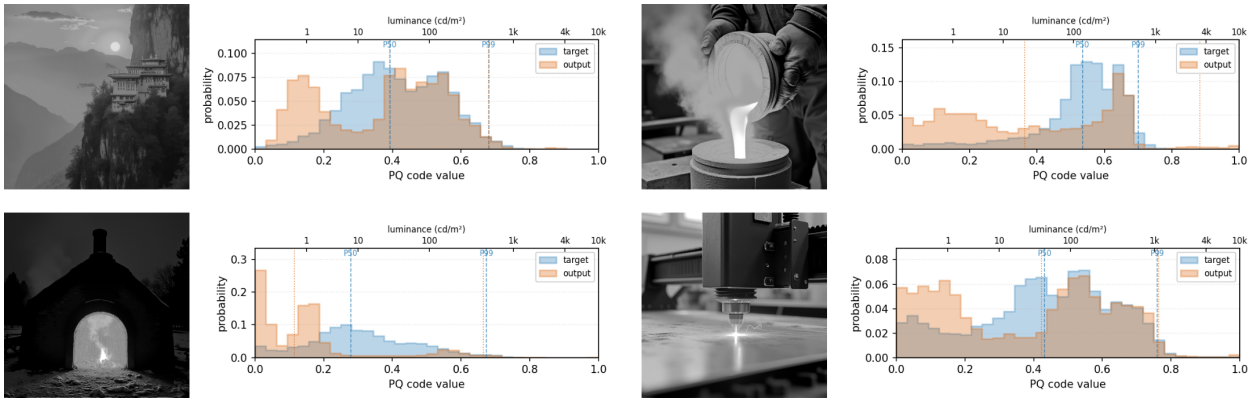


Figure 17 Failure case involving target distribution infeasibility.

gradients near saturation.

Target Distribution Infeasibility. We also observe cases, shown in Figure 17, where the generated images cannot fully match the target histogram, particularly for prompts corresponding to low-dynamic-range scenes (e.g., indoor or low-contrast environments). In such cases, the model converges to a compromise solution that only partially aligns with the target distribution. This reflects a feasibility limitation: the pretrained diffusion model defines a manifold of plausible images, and not all luminance distributions are achievable within this space. When the target distribution is incompatible with the scene semantics, the guidance objective becomes partially infeasible, resulting in persistent distribution mismatch. This suggests that effective distribution shaping requires compatibility between the target statistics and the underlying generative prior, and motivates adaptive or content-aware target specification.

L Future Work

Our analysis highlights several directions for improving distribution-level guidance.

First, saturation-induced failures suggest the need for HDR-aware decoding mechanisms that preserve gradients near extreme luminance values. Addressing this limitation is critical for accurately controlling high-intensity regions.

Second, the effectiveness of distribution shaping depends on the compatibility between the target distribution and the generative prior. Future work may explore adaptive or content-aware target specification to ensure feasibility during sampling.

Finally, we observe that existing evaluation metrics for HDR generation remain limited. Most perceptual metrics are designed for SDR content or are insensitive to luminance distribution, making them poorly aligned with HDR characteristics. Developing metrics that capture both perceptual quality and distributional fidelity in HDR space is an important direction for future research.

M Broader Impacts and Safeguards

Broader Impacts. This work improves controllable HDR image and video generation through training-free luminance distribution shaping. Potential positive impacts include lowering the computational cost of HDR content creation and enabling more accessible HDR workflows for creative applications such as digital media, film, and immersive visualization.

At the same time, improved HDR realism may increase risks already associated with generative media, including deceptive or misleading synthetic content. Since HDR imagery can produce more realistic lighting and highlight behavior, generated outputs may appear more visually convincing. However, our method does not introduce a new generative backbone or new semantic generation capabilities; it operates only as a test-time guidance mechanism on existing diffusion models.

Safeguards. Our work does not release a new large-scale generative model. The proposed method is an inference-time guidance technique applied to publicly available pretrained backbones. The lightweight text-to-histogram regressor predicts only coarse luminance distributions and cannot independently generate images.

We do not release any private user metadata or identity-related supervision. The proposed method focuses solely on luminance distribution control and does not introduce mechanisms for identity imitation, biometric analysis, or targeted manipulation.