LUMAGUIDE: DISTRIBUTION SHAPING FOR TRAINING -FREE HDR GENERATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

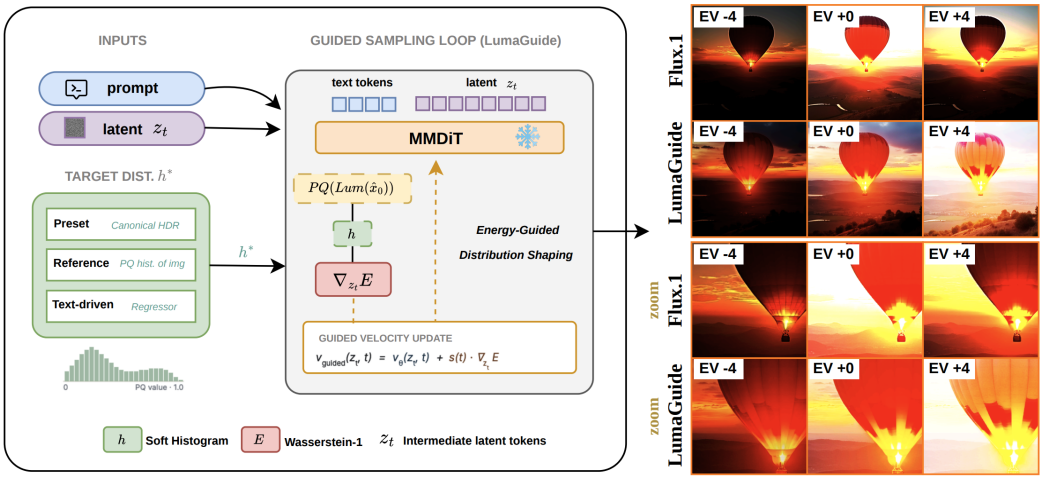

Figure 1: **LumaGuide** steers a pretrained diffusion model toward HDR-consistent luminance distributions at sampling time, without modifying any model weights. Generated outputs preserve semantic content (top row) while producing structured highlights and faithful shadow detail under exposure adjustment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

ABSTRACT

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.

1 INTRODUCTION

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)

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

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 & Nichol (2021); Ho & 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.

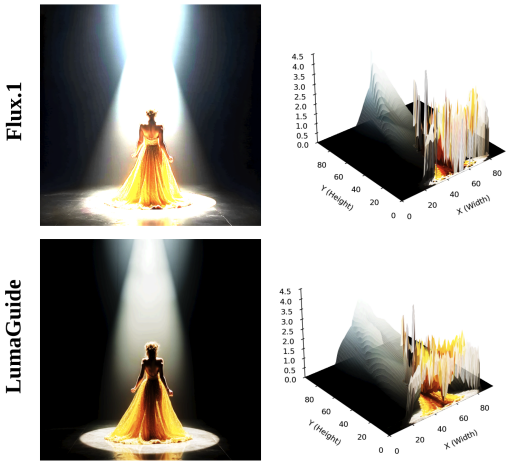

Figure 2: **Per-pixel PQ-luminance surface for an identical seed and prompt.** The Flux.1 Black Forest Labs (2024) baseline (top) overshoots into clipped highlights; LumaGuide (bottom) redistributes mass into mid-tones and preserves the highlight gradient, matching the target HDR distribution.

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 equation 2 differs sharply from posterior sampling or classifier guidance Chung et al. (2022); Ho & 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 & Ye (2021) $\hat{x}_{00}(z_t) := \mathrm{Dec}\big(z_t - t\, v_\theta(z_t, t)\big)$. The guided velocity is

$$v_{\mathrm{g}}(z_t, t) \;=\; v_\theta(z_t, t) \;+\; s(t)\, \nabla_{z_t} E(\hat{x}_{00}(z_t)), \tag{5}$$

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 $z \mapsto E(\hat{x}_{00}(z))$ 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 $\Delta E(z_t) := E(\hat{x}_{00}(z_{t+\Delta t})) - E(\hat{x}_{00}(z_t))$ 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 A.2.6.

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 A.2.7 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 $n = 0, \dots, T-1$ **do**
  $\hat{x}_{00} \leftarrow \mathrm{Dec}\big(z_{t_n} - t_n\, v_\theta(z_{t_n}, t_n)\big)$ ▷ flow decode ($\hat{x}_{00}$ is RGB)
  $Y \leftarrow \mathrm{PQ}(\mathrm{Lum}(\hat{x}_{00}))$ ▷ Per-pixel PQ luminance
  $h_\sigma \leftarrow$ soft-histogram of Y via Eq. equation 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 $\mathrm{PQ} \circ \mathrm{Lum}$
  $z_{t_{n+1}} \leftarrow z_{t_n} + (t_{n+1} - t_n)\big(v_\theta(z_{t_n}, t_n) + s(t_n)\, g_t\big)$
end for
return $\hat{x}_{00}(z_{t_T})$

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 A.2.8 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.

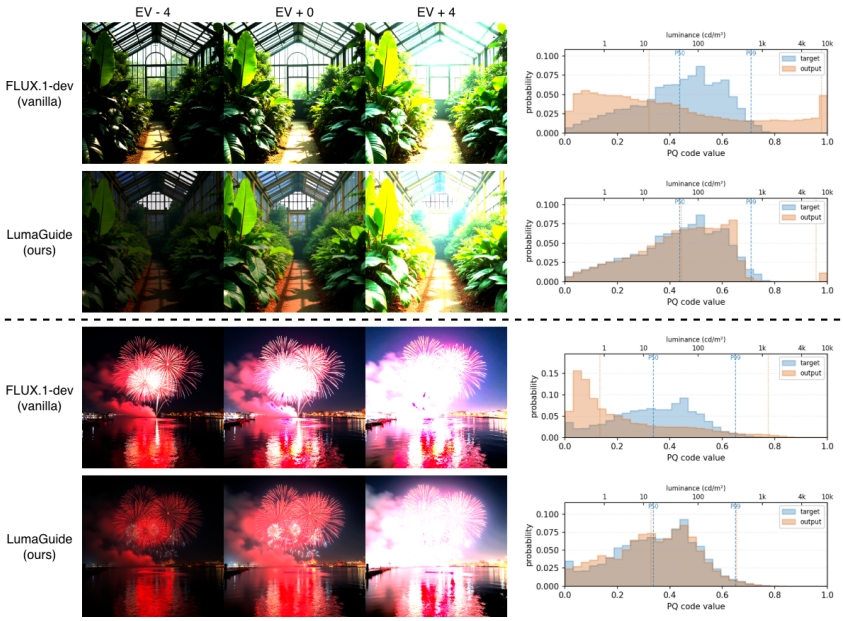

Figure 3: **HDR visualization under different exposure levels.** In PQ space, Flux.1 Black Forest Labs (2024) often yields over-exposed, non-HDR luminance distributions. **LumaGuide** steers the luminance distribution toward a target HDR distribution while preserving scene structure.

3.3 Guidance Schedule

Following previous methods Dhariwal & Nichol (2021); Ho & 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 A.4 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

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

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, $h^* = h_\sigma(x_{\mathrm{ref}})$; (c) text-driven, $h^* = \mathrm{MLP} \circ \mathrm{CLIP}(c)$ 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 A.6 and A.9. **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 A.7.

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 A.10. 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 midtones, 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 A.5.2. We select $s_0 = 2000$ as the default trade-off point.

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

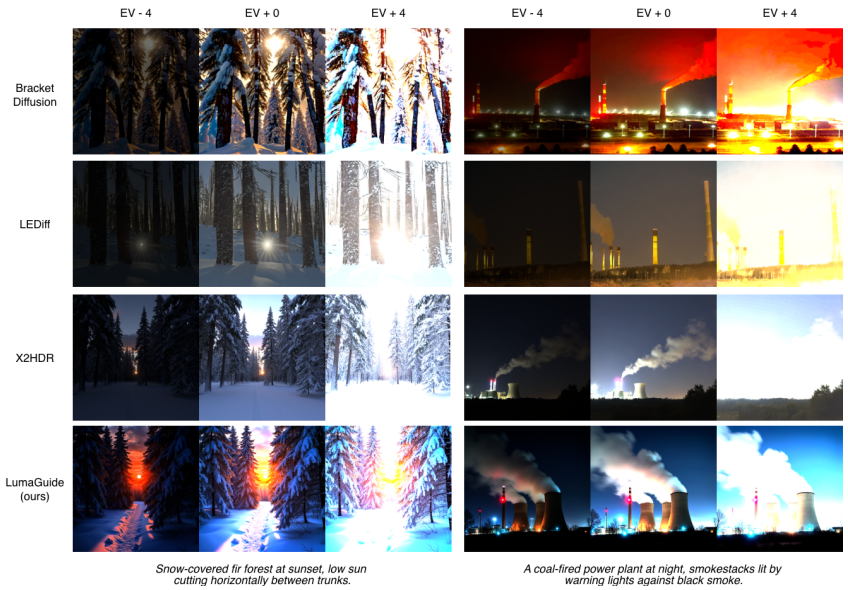

Figure 4: Qualitative comparison with existing HDR generation methods. For fair comparison, we normalize the exposure of all methods by adjusting the EV+0 images such that the median luminance is fixed at 8 nits since several existing methods are not calibrated to absolute luminance values, making direct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

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 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 A.5.

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 4 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 $p99_{dist}$ among weighting strategies, together with the highest

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

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 4: **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.

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 A.7.

Table 5: **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 retraining. Table 5 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 A.6 and A.9.

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.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation and debugging of research code and to discuss and refine research ideas conceived by the authors. We have not used generative AI tools to originate the core research idea or to determine the scientific conclusions of this work, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to improve the clarity, grammar, and presentation of the manuscript. We have reviewed all AI-assisted work. AI-assisted code was inspected, modified as needed, and tested by the authors for correctness. Suggestions arising from discussions with generative AI were critically evaluated by the authors and validated through our own analysis and experiments. All AI-assisted manuscript revisions were reviewed to ensure technical accuracy and consistency with the authors' intended claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves a human-subject study conducted to evaluate the generated HDR image quality. All participants provided informed consent prior to participation and were informed of the study procedures and their right to withdraw. We collected only the information necessary for conducting the study and analyzing the experimental results. Participant data were handled in accordance with the approved protocol and were anonymized or de-identified before analysis and reporting. No personally identifiable information is included in the paper or released with the research data. The proposed method and the associated human-subject evaluation are intended for research purposes. We have taken appropriate measures to protect participant privacy and to report the study results accurately and responsibly.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. We report the model configurations, hyperparameters, evaluation metrics, and baseline settings used in our experiments to enable consistent comparison. We will also provide the implementation and configuration files associated with the experiments to facilitate reproduction of the reported results.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 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)*, pp. 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*, pp. 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, pp. 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*, pp. 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*, pp. 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)*, pp. 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, pp. 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*, pp. 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*, pp. 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)*, pp. 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)*, pp. 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*, pp. 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*, pp. 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*, pp. 10621–10631, 2025.