RATIODIFF: ALIGNING DIFFUSION MODELS VIA PREFERENCE RATIO MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved remarkable success in text-to-image generation, producing diverse and visually compelling images from natural language prompts. Preference learning further aligns these models with human judgments of image quality and prompt adherence. However, effective preference alignment faces two challenges: weak corrective signals on strongly misranked pairs and difficulty balancing preference improvement with denoising preservation. We introduce RatioDiff, an offline framework centered on ratio diffusion preference learning. Our core contribution is a preference ratio formulation that explicitly accounts for distinct noisy diffusion states through a reference correction. This formulation leads to a tractable Bregman objective that sustains corrective pressure on misranked pairs, providing a principled foundation for diffusion preference optimization. To support this objective, we introduce learned pair weights guided by reference consistency, separating the strength of preference correction from each pair's contribution to training. We further develop adaptive regularization that controls reference proximity and preferred-image denoising relative to the preference update, limiting model drift while preserving denoising behavior. Together, these contributions combine preference ratio matching with explicit control over pair influence and regularization, without additional inference overhead. Experiments on multiple text-to-image backbones and preference benchmarks show that RatioDiff consistently improves human-preference alignment while preserving and enhancing visual quality and text–image consistency.

1 INTRODUCTION

Diffusion-based text-to-image (T2I) models generate high-quality and diverse images from natural language prompts (Ho et al., 2020; Rombach et al., 2022; Podell et al., 2024). However, denoising large-scale image–text data does not directly capture human preferences. Generated images may still violate requested relationships, contain artifacts, or deviate from the intended style. Human preference feedback provides supervision for these distinctions and motivates alignment beyond the pretraining objective (Kirstain et al., 2023; Xu et al., 2023).

To incorporate such feedback into diffusion models, existing alignment methods broadly follow reinforcement-learning and direct preference-learning paradigms, either optimizing reward signals or learning directly from pairwise human comparisons. Reinforcement learning from human feedback (RLHF) aligns generative models by learning a reward model from human judgments and optimizing the model against that reward (Christiano et al., 2017; Ouyang et al., 2022). In T2I generation, reward-based fine-tuning and diffusion policy optimization improve outputs through feedback (Lee et al., 2023; Black et al., 2024; Fan et al., 2023), but typically require online generation and reward evaluation. Direct Preference Optimization (DPO) instead learns directly from offline comparisons through policy-to-reference likelihood ratios (Rafailov et al., 2023). Diffusion-DPO adapts this principle to diffusion models using a tractable denoising surrogate (Wallace et al., 2024), followed by extensions based on utility optimization, divergence minimization, score matching, and margin-based learning (Li et al., 2024; 2025; Zhu et al., 2025; Hong et al., 2026).

Despite this progress, two challenges remain. First, *gradient attenuation limits correction on strongly misranked pairs*. Nonlinear utility or divergence objectives can assign vanishingly small gradients to comparisons with a large negative preferred-versus-rejected margin (Li et al., 2024;

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.

2025). While these approaches improve robustness to noisy feedback, gradient attenuation also suppresses updates on correctly annotated pairs that the model ranks in the opposite order with high confidence. Effective alignment should therefore sustain correction on such pairs while controlling each pair's contribution separately. Second, *preference improvement should be balanced with denoising preservation*. Existing approaches combine preference learning with denoising objectives to retain generative quality (Zhu et al., 2025; Hong et al., 2026). However, a fixed loss coefficient does not control the ratio between auxiliary and preference gradient norms as these norms vary across samples, noise levels, and training stages. Excessive regularization suppresses preference updates, whereas insufficient regularization leaves model drift weakly constrained. This motivates adapting the regularization strength to the current preference gradient.

To address these challenges, we introduce **RatioDiff**, an offline framework for *ratio diffusion preference learning*. We formulate preference learning directly over pairs of denoising transitions and derive a reference correction for the different noisy states on which they are conditioned. This correction accounts for state-dependent normalization in the preference ratio. Under Gaussian transitions, the ratio admits a tractable noise-prediction margin, which we optimize with a Bregman objective that sustains corrective pressure on misranked pairs. Two complementary components control how this objective updates the denoiser. A trainable confidence head learns reference-consistency weights to adjust each noisy pair's contribution independently of its corrective signal. An adaptive regularization scheme controls reference proximity and preferred-image denoising by bounding auxiliary gradients relative to the corresponding preference gradient in denoiser-output space. These bounds use analytic gradient norms and require no additional denoiser evaluations. At inference time, only the fine-tuned diffusion model is retained.

Our contributions are threefold: (i) **Diffusion preference ratio matching.** We derive a reference-corrected preference-ratio formulation with exact ratio consistency under explicit assumptions, together with a tractable Gaussian surrogate that maintains correction on misranked pairs. (ii) **Adaptive weighting and regularization.** Learned pair weights and auxiliary gradient budgets explicitly control pair influence and denoising preservation without inference overhead. (iii) **Empirical evaluation.** Experiments on three prompt benchmarks demonstrate improvements over the pretrained model in preference scores, aesthetic quality, and text–image alignment.

2 Related Work

Text-to-image diffusion models. Diffusion models learn to reverse a stochastic corruption process that transforms data into noise (Sohl-Dickstein et al., 2015; Ho et al., 2020). Score-based modeling offers a related view through gradients of noisy data distributions (Song & Ermon, 2019), unified with diffusion models via stochastic differential equations by Song et al. (2021). Karras et al. (2022) further systematize diffusion design choices for improved quality and efficiency. Building on these foundations, latent diffusion and SDXL enable high-resolution text-conditioned generation (Rombach et al., 2022; Podell et al., 2024).

Alignment of T2I models with human preferences. RLHF learns rewards from human judgments and optimizes the model accordingly (Christiano et al., 2017; Ouyang et al., 2022). For T2I models, reward-based methods include reward-weighted fine-tuning (Lee et al., 2023), DDPO (Black et al., 2024), DPOK (Fan et al., 2023), ReFL (Xu et al., 2023), and differentiable-reward optimization via DRaFT and AlignProp (Clark et al., 2024; Prabhudesai et al., 2023). DPO instead learns directly from pairwise comparisons without explicit reward modeling or online RL (Rafailov et al., 2023), and Diffusion-DPO adapts this idea with a denoising surrogate (Wallace et al., 2024). Later approaches explore utility optimization (Li et al., 2024), reverse-KL minimization (Li et al., 2025), preference-aware score matching (Zhu et al., 2025), and reference-free margin learning with preferred-image denoising (Hong et al., 2026). RatioDiff focuses on preference ratio matching with reference-consistency weighting and adaptive regularization; its preferred-image auxiliary is related to DSPO, but its gradient is explicitly controlled relative to the preference gradient.

Bregman ratio matching. Bregman divergences provide a general framework for density-ratio estimation (Sugiyama et al., 2012). BPO applies this principle to preference learning with a scaled Basu generator (Kim et al., 2025), while TokenRatio extends it to token-level decisions with state-dependent corrections (Nguyen et al., 2026). RatioDiff extends these ideas to continuous diffusion

transitions through reference-corrected ratios across distinct noisy states. Under Gaussian transitions, this reduces to a tractable noise-prediction margin and yields a Bregman objective that maintains corrective pressure on misranked pairs.

3 PRELIMINARIES

Signal-to-noise diffusion models. Let c denote the text condition and $x_0 \in \mathbb{R}^d$ an image or scaled latent representation (Rombach et al., 2022). For schedules α_t, σ_t, the forward and reverse processes are

$$\begin{aligned} q(x_t \mid x_0) &= \mathcal{N}(\alpha_t x_0, \sigma_t^2 I), \qquad & x_t &= \alpha_t x_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \\ p_\theta(x_{0:T} \mid c) &= p_T(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, c), \quad & p_T &= \mathcal{N}(0, I). \end{aligned} \tag{1}$$

The signal-to-noise ratio $\lambda_t = \alpha_t^2/\sigma_t^2$ decreases with t, and a variance-preserving schedule satisfies $\alpha_t^2 + \sigma_t^2 = 1$. With $\nu_t^2 = \sigma_t^2 - (\alpha_t/\alpha_{t-1})^2 \sigma_{t-1}^2$, stochastic reverse steps take the form (Ho et al., 2020; Kingma et al., 2021):

$$\begin{aligned} p_\theta(x_{t-1} \mid x_t, c) &= \mathcal{N}(\mu_\theta(x_t, t, c), v_t I), \qquad & v_t &= \frac{\nu_t^2 \sigma_{t-1}^2}{\sigma_t^2}, \\ \mu_\theta(x_t, t, c) &= \frac{\alpha_{t-1}}{\alpha_t} x_t - b_t \epsilon_\theta(x_t, t, c), \quad & b_t &= \frac{\alpha_{t-1} \nu_t^2}{\alpha_t \sigma_t}. \end{aligned} \tag{2}$$

The posterior $q(x_{t-1} \mid x_t, x_0)$ uses the same covariance, with ϵ replacing ϵ_θ in the mean. Pretraining minimizes $\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{c,x_0,t,\epsilon}[\omega_{\mathrm{train}}(t) \|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2]$. For transition ratios, we sample $t \sim \mathcal{U}(\mathcal{T})$ with $\mathcal{T} = \{t : v_t > 0,\ \alpha_t \sigma_t > 0\}$, excluding deterministic endpoints.

Direct preference optimization. An offline dataset $\mathcal{D} = \{(c, x_0^w, x_0^\ell) : x_0^w \succ x_0^\ell\}$ records preferred and rejected images. The Bradley–Terry model (Bradley & Terry, 1952) specifies $p(x_0^w \succ x_0^\ell \mid c) = \mathrm{sigm}(r(c, x_0^w) - r(c, x_0^\ell))$, where $\mathrm{sigm}(u) = (1 + e^{-u})^{-1}$ is the sigmoid function. Maximizing $\mathbb{E}_p[r(c, x_0)] - \beta D_{\mathrm{KL}}(p \| p_{\mathrm{ref}})$ gives $p^*(x_0 \mid c) = p_{\mathrm{ref}}(x_0 \mid c) e^{r(c,x_0)/\beta}/Z(c)$, for $\beta > 0$ and a finite normalizer $Z(c) = \sum_{x_0} p_{\mathrm{ref}}(x_0 \mid c) \exp\left(\frac{r(c,x_0)}{\beta}\right)$. DPO substitutes this reward–policy relationship into the comparison likelihood (Rafailov et al., 2023):

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\mathbb{E}_{\mathcal{D}}\left[\log \mathrm{sigm}\left[\beta\left(\ell_\theta(x_0^w, c) - \ell_\theta(x_0^\ell, c)\right)\right]\right], \qquad \ell_\theta(x_0, c) = \log \frac{p_\theta(x_0 \mid c)}{p_{\mathrm{ref}}(x_0 \mid c)}. \tag{3}$$

For diffusion models, the marginal $p_\theta(x_0 \mid c)$ is generally intractable; Diffusion-DPO therefore constructs a tractable denoising surrogate (Wallace et al., 2024).

Density-ratio matching. For positive densities p_+, p_-, define $R_* = p_-/p_+$. Given a strictly convex, twice continuously differentiable generator h, let $B_h(u\|v) = h(u) - h(v) - h'(v)(u - v)$. Then

$$\begin{aligned} D_h(R_* \| R_\theta) &= \mathbb{E}_{p_+}\left[B_h(R_* \| R_\theta)\right] \\ &= \mathbb{E}_{p_+}[h'(R_\theta) R_\theta - h(R_\theta)] - \mathbb{E}_{p_-}\left[h'(R_\theta)\right] + C, \end{aligned} \tag{4}$$

where C is independent of θ. Under realizability and finite expectations, the minimum divergence is zero exactly when $R_\theta = R_*$ almost everywhere. This yields ratio learning from samples without estimating the two densities separately (Sugiyama et al., 2012; Kim et al., 2025).

4 METHODOLOGY

We introduce RatioDiff, which aligns diffusion models with human preferences by matching preference ratios at each denoising step, learning how much each noisy pair contributes, and regularizing updates to preserve denoising behavior.

4.1 Reinforcement Learning View of Denoising

We view denoising as a sequence of decisions. Given a sample image $x_0 \sim q(x_0)$, we use the forward process described in Section 3 to generate the trajectory $x_{0:T}$, where each $x_t \sim q(x_t \mid x_0)$ for $t = 1, \ldots, T$ (i.e., $x_t = \alpha_t x_0 + \sigma_t \epsilon$). We denote $y_t = x_{T-t}$ for $t = 0, \ldots, T$. Let the state $s_t = [c, y_{<t}]$ and the action $a_t = y_t$. We then consider the image-based reward function $R_t = R(s_t, a_t) = r([c, y_{<t}], y_t)$, which can be interpreted as follows: given the prompt c and the partially denoised image $y_{t-1} = x_{T-t+1}$, we assign a high reward to $y_t = x_{T-t}$ when it exhibits high fidelity or quality relative to the original image $x_0 = y_T$. Using these definitions, for a policy p[1] we define the state–action value function Q^p, the state value function V^p, and the advantage function A^p as

$$\begin{aligned} Q^p([c, y_{<t}], y_t) &= \mathbb{E}_p \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, s_t = [c, y_{<t}],\ a_t = y_t \right], \\ V^p([c, y_{<t}]) &= \mathbb{E}_p[Q^p([c, y_{<t}], y_t) \,|\, s_t = [c, y_{<t}]], \\ A^p([c, y_{<t}], y_t) &= Q^p([c, y_{<t}], y_t) - V^p([c, y_{<t}]), \end{aligned} \tag{5}$$

where $\gamma \in [0, 1]$ is the discount factor: smaller values give less weight to rewards from later denoising steps.

4.2 Ratio Diffusion Preference Learning

Using the denoising policy and reference advantage from Section 4.1, we learn the backward distribution $p_\theta(y_t \mid y_{<t}, c) = p_\theta(y_t \mid y_{t-1}, c)$. At the token level, we aim to learn the optimal backward probability as follows:

$$\max_{p_\theta} \mathbb{E}_{c, y_{<t}} \left[\mathbb{E}_{y_t \sim p_\theta(\cdot \mid y_{t-1}, c)} [A_{\text{ref}}([c, y_{<t}], y_t)] - \beta D_{\text{KL}}(p_\theta(\cdot \mid c, y_{<t}) \,\|\, p_{\text{ref}}(\cdot \mid c, y_{<t})) \right]. \tag{6}$$

The outer expectation uses a fixed distribution of prompts and histories, and β plays the same regularization role as in Equation 3. For a pair $(c, x_0^w, x_0^l) \in \mathcal{D}$, we follow the advantage-based construction of TokenRatio (Nguyen et al., 2026) and write

$$\begin{aligned} p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l) &= \text{sigm}\big(A_{\text{ref}}([c, y_{<t}^w], y_t^w) - A_{\text{ref}}([c, y_{<t}^l], y_t^l)\big), \\ p_{\text{data}}(y^w \succ y^l \mid c) &= \prod_t p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l). \end{aligned} \tag{7}$$

Here, $y^w = y_{0:T}^w$ and $y^l = y_{0:T}^l$ denote the paired trajectories, and p_{data} characterizes their preference relation. The trajectory-level preference probability is obtained by factorizing over the corresponding local comparisons along the trajectories. This factorization, together with the sampling conditions under which it is used, is detailed in Appendix A.3.

Proposition 1. *Let π_{θ^*} be the solution of (6). Then*

$$\frac{p_{\theta^*}(y_t^w \mid c, y_{t-1}^w)}{p_{\theta^*}(y_t^l \mid c, y_{t-1}^l)} = \frac{p_{\text{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\text{ref}}(y_t^l \mid c, y_{t-1}^l)} \left(\frac{p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\text{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)} \right)^{\frac{1}{\beta}} \omega_t, \tag{8}$$

where $\omega_t = \exp\{D_{\text{KL}}(p_{\text{ref}}(\cdot \mid y_{t-1}^l, c) \| p_{\theta^*}(\cdot \mid y_{t-1}^l, c)) - D_{\text{KL}}(p_{\text{ref}}(\cdot \mid y_{t-1}^w, c) \| p_{\theta^*}(\cdot \mid y_{t-1}^w, c))\}$ accounts for the different normalization factors at the two noisy states. The assumptions and derivation are given in Appendix A.1. Inspired by Proposition 1, we reformulate preference optimization as matching the data ratio $R_{\text{data}}^t(y_t^w, y_t^l \mid y_{<t}^w, y_{<t}^l, c)$ and the model ratio $R_\theta^t(y_t^w, y_t^l \mid y_{<t}^w, y_{<t}^l, c)$:

$$R_{\text{data}}^t = \frac{p_{\text{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}, \tag{9}$$

$$R_\theta^t = \left(\frac{p_\theta(y_t^l \mid c, y_{t-1}^l)\, p_{\text{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\text{ref}}(y_t^l \mid c, y_{t-1}^l)\, p_\theta(y_t^w \mid c, y_{t-1}^w)} \, \omega_t \right)^\beta. \tag{10}$$

[1] The policy p here is the backward transition probability $p(x_{t-1} \mid x_t, c)$.

We suppress the conditioning arguments below. These ratios express rejected-versus-preferred odds, so a value below one favors the preferred branch. Proposition 1 gives $R^t_{\theta^*} = R^t_{\text{data}}$. For a trainable ratio, ω_t is evaluated with the current p_θ in both KL terms; this recovers the stated correction at the optimum. Using Bregman ratio matching (Sugiyama et al., 2012), we define the loss at step t as

$$D^t_h(R^t_{\text{data}}, R^t_\theta) = \mathbb{E}_{p_{\text{data}}(y^w_{\le t} \succ y^l_{\le t}|c)}\Big[h(R^t_{\text{data}}) - h(R^t_\theta) - h'(R^t_\theta)(R^t_{\text{data}} - R^t_\theta)\Big]. \tag{11}$$

We use the generator and Bregman divergence defined in Equation 4. The expectation is over ordered preferred–rejected trajectory prefixes, with averaging over prompts understood. We keep the objective at a fixed step t; the sampling and learned weighting are introduced in Section 4.3.

Proposition 2. *For a fixed step t, under realizability and the comparison-support conditions, $p_{\theta^*}(\cdot \mid c, y_{t-1}) = p_{\widehat{\theta}}(\cdot \mid c, y_{t-1})$ on the covered support at this step, where*

$$\pi_{\widehat{\theta}} = \arg\min_{\pi_\theta} D^t_h(R^t_{\text{data}}, R^t_\theta). \tag{12}$$

Appendix A.2 gives the support conditions and proof. Although ratio matching recovers the local optimum, its loss contains unknown data ratios. The next result expresses it using only model ratios, making it evaluable from preference pairs.

Proposition 3. *Under the ordered-comparison sampling conditions, we have $\mathcal{L}_t(\theta, p_{\text{data}}) = D^t_h(R^t_{\text{data}}, R^t_\theta) + \text{const}$, where*:

$$\mathcal{L}_t(\theta, p_{\text{data}}) = \mathbb{E}_{p_{\text{data}}(y^w \succ y^l|c)}\left[h'(R^t_\theta)R^t_\theta - h(R^t_\theta) - h'\big((R^t_\theta)^{-1}\big)\right]. \tag{13}$$

The constant is independent of θ. The loss in Equation 13 depends only on the ratio R^t_θ; different Bregman ratio estimation methods differ in their choice of the convex generator h. Following BPO (Kim et al., 2025), we use the scaled Basu power-divergence generator

$$h(R) = \frac{R^{1+\lambda} - R}{s_B\lambda(\lambda+1)}, \qquad R > 0, \quad \lambda > 0, \quad s_B > 0. \tag{14}$$

The shape parameter λ adjusts the loss response across ratio values, including confidently ordered pairs ($R \ll 1$) and nearly tied pairs ($R \approx 1$), while s_B sets the overall scale. Appendix A.5 gives the derivatives and the limit at $\lambda = 0$; the proof of Proposition 3 and its sampling conditions are in Appendix A.3. It remains to estimate R^t_θ in Equation 10. We use the Gaussian posterior of the forward process. For forward timesteps $s < t$, where s here is a time index, this posterior is

$$q(x_s \mid x_t, x_0) = \mathcal{N}\Big(m_{s|t}, \sigma^2_{s|t} I\Big), \tag{15}$$

with mean and scalar variance

$$\begin{aligned} m_{s|t} &= \frac{\alpha_t\sigma_s^2}{\alpha_s\sigma_t^2}x_t + \frac{\alpha_s^2\sigma_t^2 - \alpha_t^2\sigma_s^2}{\alpha_s\sigma_t^2}x_0, \\ \sigma^2_{s|t} &= \sigma_s^2\left(1 - \frac{\alpha_t^2\sigma_s^2}{\alpha_s^2\sigma_t^2}\right). \end{aligned} \tag{16}$$

The reverse model replaces the unknown clean image in this posterior with its denoising estimate:

$$\begin{aligned} p_\theta(x_s \mid x_t, c) &= q(x_s \mid x_t, \widehat{x}_{0,\theta}(x_t, t, c)), \\ \widehat{x}_{0,\theta}(x_t, t, c) &= \frac{x_t - \sigma_t\epsilon_\theta(x_t, t, c)}{\alpha_t}. \end{aligned} \tag{17}$$

The reference uses the same parameterization with its frozen noise prediction. For adjacent steps $s = t - 1$, the variance is v_t from Equation 2. These Gaussian densities make the likelihood and KL terms in the ratio tractable. We denote the resulting Gaussian surrogate ratio by $R_{\theta,t}$, where t indexes the forward noise level. Its expression in terms of denoiser outputs and its relation to the exact transition ratio are provided in Appendix A.4.

4.3 Learning the Ratio Model

Motivated by the observation of Hang et al. (2023) that different noise levels can lead to distinct learning behaviors, we account for the varying strength of preference signals across denoising timesteps. Accordingly, we modulate the preference update using two complementary coefficients: a base update scale S_R and a learned pair-specific weight $\text{sg}(w_t)$.

Controlling the update strength. The ratio temperature β controls how sensitively $R_{\theta,t} = \exp(\beta m_{\theta,t})$ responds to the denoising margin, but also scales its gradient. Here $s_B > 0$ is the fixed scale parameter of the Basu generator in Equation 14: both h and the comparison loss ℓ_h are proportional to $1/s_B$. In particular, the unscaled loss has margin derivative $2\beta/s_B$ at $m_{\theta,t} = 0$ ($R_{\theta,t} = 1$). We introduce a desired initial update strength $S_R > 0$ and use the calibrated multiplier

$$s_R = \frac{s_B S_R}{2\beta}. \tag{18}$$

The coefficient s_R multiplies the comparison loss. Its factor s_B compensates for the $1/s_B$ scale of the Basu generator, so the calibrated loss has initial margin slope S_R independently of s_B and β. Away from initialization, β still controls the nonlinear response of the ratio. Appendix A.5 gives the derivation.

Learning how much each noisy pair contributes. Human preferences compare clean images, but their ordering can be ambiguous after adding noise. We therefore learn how much each noisy pair should contribute. A confidence head g_ϕ predicts whether the frozen reference denoises the preferred branch more accurately:

$$p_{\phi,t} = \mathrm{sigm}(g_\phi(\mathrm{sg}(Z_t))), \qquad p_{\phi,t} \in (0,1). \tag{19}$$

The detached features Z_t summarize the noisy pair, reference predictions, and timestep. Since no human labels are available at individual noise levels, we form a proxy target from reference errors:

$$E^*_{\mathrm{ref},t} = \|\epsilon^* - \epsilon^*_{\mathrm{ref}}\|_2^2, \qquad * \in \{w, l\}. \tag{20}$$

We set $d_t = 1$ when $E^w_{\mathrm{ref},t} < E^l_{\mathrm{ref},t}$, $d_t = 0$ when the inequality is reversed, and $d_t = 1/2$ for a tie. With optional smoothing $\widetilde{d}_t = (1-\eta)d_t + \eta/2$ ($0 \le \eta < 1$), the head is trained by

$$\mathcal{L}_{\mathrm{conf}}(\phi) = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\ t\sim\mathcal{U}(\mathcal{T}),\\ \epsilon^w,\epsilon^l \overset{\mathrm{iid}}{\sim}\mathcal{N}(0,I)}} \left[-\widetilde{d}_t \log p_{\phi,t} - (1-\widetilde{d}_t)\log(1-p_{\phi,t}) \right]. \tag{21}$$

This teaches g_ϕ reference consistency rather than the unobserved human preference at each noise level. We convert its prediction to a positive weight and normalize across mini-batch $\mathcal{B}$:

$$\begin{aligned} \widetilde{w}_i &= w_{\min} + (1 - w_{\min}) p_{\phi,i}, \\ w_i &= \frac{\widetilde{w}_i}{|\mathcal{B}|^{-1} \sum_{k\in\mathcal{B}} \widetilde{w}_k}, \qquad 0 < w_{\min} < 1. \end{aligned} \tag{22}$$

Here i indexes a pair at its sampled timestep. The floor $w_{\min}$ retains every pair, while normalization gives unit mean weight. We detach the normalized weight w_t in the preference loss, so g_ϕ learns only from Equation 21. Combining this weight with the calibrated comparison loss gives

$$\mathcal{L}_{\mathrm{pref},t} = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\ t\sim\mathcal{U}(\mathcal{T}),\\ \epsilon^w,\epsilon^l \overset{\mathrm{iid}}{\sim}\mathcal{N}(0,I)}} \left[\mathrm{sg}(w_t) \frac{s_B S_R}{2\beta} \left(h'(R_{\theta,t}) R_{\theta,t} - h(R_{\theta,t}) - h'(R_{\theta,t}^{-1}) \right) \right]. \tag{23}$$

Both losses use the same sampled pair and timestep, with independent noise for the two branches; w_t also depends on its normalization context in Equation 22. This evaluates the per-comparison loss from Equation 13 at the Gaussian ratio and scales it by s_R. Details of the head and its proxy target appear in Appendix A.6.

4.4 Regularized Preference Fine-Tuning

The ratio objective improves pairwise preference by increasing the relative likelihood of the preferred transition over the rejected one. However, this relative supervision has two important limitations: it does not explicitly control how far either branch deviates from the pretrained denoiser (Gheshlaghi Azar et al., 2024; Fan et al., 2023), and it may improve the preference margin simply by degrading the rejected transition rather than by enhancing denoising of the preferred image (Xu et al., 2025). To address these issues and promote more stable optimization, we introduce two complementary regularizers. For $* \in \{w, l\}$, write $\epsilon^*_\theta = \epsilon_\theta(x^*_t, t, c)$, and let $\epsilon^*_{\mathrm{ref}}$ denote the frozen

reference prediction on the same input. Reference drift is controlled by anchoring both branches to the frozen pretrained denoiser:

$$\mathcal{L}_{\mathrm{ref},t} = \chi_t \sum_{*\in\{w,l\}} \|\epsilon_\theta^* - \epsilon_{\mathrm{ref}}^*\|_2^2, \qquad \chi_t = \frac{b_t^2}{2v_t}, \tag{24}$$

where b_t and v_t are defined in Equation 2. Under the shared Gaussian covariance, this is the sum of the two conditional transition KL divergences, penalizing reference drift even in directions not captured by the ratio margin. We further apply direct denoising supervision to the preferred image, providing an absolute target beyond the pairwise objective. Let ϵ^w denote the noise used to corrupt the preferred image; we define:

$$\mathcal{L}_{\mathrm{win},t} = \left\|(\epsilon_\theta^w - \epsilon^w) - \kappa_t(\epsilon_\theta^w - \epsilon_{\mathrm{ref}}^w)\right\|_2^2, \tag{25}$$

in which $\kappa_t \in [0, \kappa_{\max}]$ controls the adaptive correction strength, with a fixed bound $0 \le \kappa_{\max} < 1$. It decreases as the preferred–rejected denoising gap improves relative to the reference. We detach κ_t, so gradients flow only through the preferred branch. Appendix A.7 provides the gate definition and resulting effective target.

To keep these safeguards from overwhelming preference learning, we adapt their coefficients for each pair and timestep. We limit the norm of each regularizer's weighted gradient to a fixed budget fraction ρ_{ref} or ρ_{win} of the preference-gradient norm, measured with respect to the same denoiser outputs. With detached coefficients $a_{\mathrm{ref},t}$ and $a_{\mathrm{win},t}$, the combined loss is

$$\mathcal{L}_{\mathrm{reg},t} = a_{\mathrm{ref},t}\mathcal{L}_{\mathrm{ref},t} + a_{\mathrm{win},t}\mathcal{L}_{\mathrm{win},t}. \tag{26}$$

The coefficients $a_{k,t}$ are computed automatically from the current gradient norms at each training step, rather than tuned as fixed loss weights. The rule scales down a regularizer whenever its gradient would exceed the prescribed budget ρ_k. Appendix A.8 gives the coefficient rule, analytic gradient norms, and the bound in denoiser-output space.

Overall training objective. We train the denoiser and confidence head with

$$\mathcal{L}_{\mathrm{RatioDiff}}(\theta,\phi) = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\ t\sim\mathcal{U}(\mathcal{T}),\\ \epsilon^w,\epsilon^l\overset{\mathrm{iid}}{\sim}\mathcal{N}(0,I)}}\Big[\mathrm{sg}(w_t)\frac{s_B S_R}{2\beta}\big(h'(R_{\theta,t})R_{\theta,t} - h(R_{\theta,t}) - h'(R_{\theta,t}^{-1})\big) + \mathcal{L}_{\mathrm{reg},t}\Big] + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}(\phi). \tag{27}$$

Both branches share the sampled timestep with independent noise. The weights w_t, gate κ_t, and coefficients $a_{k,t}$ are detached during differentiation. Thus, the denoiser is updated by the preference and regularization losses, while the confidence head is trained only by $\lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}$. The reference model and image/text encoders remain frozen, and inference uses only the fine-tuned diffusion model. Appendix B provides further training details.

5 Experiments

5.1 Experimental Setup

We evaluate Stable Diffusion v1.5 (SD1.5) (Rombach et al., 2022) and Stable Diffusion XL 1.0 (SDXL) (Podell et al., 2024), trained on the 851,293 preference pairs in Pick-a-Pic v2 (Kirstain et al., 2023). Baselines include each pretrained model, supervised fine-tuning (SFT), and Diffusion-DPO (Wallace et al., 2024); SD1.5 additionally uses Diffusion-KTO (Li et al., 2024), DMPO (Li et al., 2025), and DSPO (Zhu et al., 2025), while SDXL includes MaPO (Hong et al., 2026). Since official DMPO code is unavailable, we reimplement it; its objective and implementation are described in Appendix C.4.

We evaluate on 500 Pick-a-Pic v2 prompts (Kirstain et al., 2023), 1,632 PartiPrompt prompts (Yu et al., 2022), and 400 HPSv2 prompts (Wu et al., 2023), reporting PickScore, HPSv2, LAION Aesthetics Score, CLIP Score, and ImageReward. All methods with the same backbone use the same prompts and generation protocol. Further evaluation and training details are provided in Appendices C.1 and C.2, respectively.

Table 1: Quantitative reward scores on SD1.5 across three validation datasets. ∗: We retrain DSPO using official code because its weights are unavailable. †: We reimplement DMPO following the original paper's settings because no official code is available. Bold and underlined values mark the best and second-best results per dataset, including ties. "Diff." denotes "Diffusion."

Dataset	Method	PickScore	HPSv2	Aesthetics	CLIP	Image Reward
Pick-a-Pic v2	SD1.5	0.2074	0.2671	5.4461	0.3325	0.1009
	SFT	**0.2131**	0.2778	5.6147	0.3430	0.6349
	Diff.-DPO	0.2119	0.2723	5.5564	0.3411	0.3070
	Diff.-KTO	0.2129	0.2780	5.6744	0.3451	0.6175
	DMPO†	0.2122	0.2705	5.6314	0.3453	0.5412
	DSPO∗	0.2124	0.2784	5.6775	0.3442	0.6242
	RatioDiff	0.2124	**0.2826**	**5.6878**	**0.3458**	**0.6772**
PartiPrompt	SD1.5	0.2151	0.2749	5.3638	0.3330	0.2308
	SFT	**0.2184**	0.2819	5.5915	0.3401	0.6455
	Diff.-DPO	0.2176	0.2783	5.4391	0.3380	0.3784
	Diff.-KTO	0.2181	0.2811	5.5678	0.3396	0.6472
	DMPO†	0.2173	0.2758	5.5438	0.3398	0.6414
	DSPO∗	0.2180	0.2791	5.5857	0.3388	0.6418
	RatioDiff	0.2178	**0.2821**	**5.5932**	**0.3402**	**0.6614**
HPSv2	SD1.5	0.2103	0.2716	5.5838	0.3545	0.1040
	SFT	0.2166	0.2828	5.8144	0.3638	0.7156
	Diff.-DPO	0.2144	0.2754	5.7042	0.3600	0.2946
	Diff.-KTO	**0.2170**	0.2832	5.8154	0.3661	0.7028
	DMPO†	0.2165	0.2668	5.7997	0.3609	0.6350
	DSPO∗	0.2169	0.2824	**5.8204**	0.3646	0.7240
	RatioDiff	0.2169	**0.2850**	5.8203	**0.3665**	**0.7398**

Table 2: Quantitative reward scores on SDXL across three validation datasets (higher is better). Bold and underlined values mark the best and second-best results per dataset, including ties. "Diff." denotes "Diffusion."

Dataset	Method	PickScore	HPSv2	Aesthetics	CLIP	Image Reward
Pick-a-Pic v2	SDXL	0.2203	0.2805	5.9618	0.3653	0.7631
	SFT	0.2171	0.2761	5.6932	0.3670	0.6526
	Diff.-DPO	**0.2256**	0.2865	6.0270	0.3742	0.8971
	MaPO	0.2212	0.2830	**6.1570**	0.3628	0.8524
	RatioDiff	0.2237	**0.2931**	6.0313	**0.3745**	**0.9232**
PartiPrompt	SDXL	0.2266	0.2839	5.6794	0.3569	0.7665
	SFT	0.2214	0.2813	5.5763	0.3581	0.7226
	Diff.-DPO	**0.2295**	0.2863	5.8068	0.3660	1.0788
	MaPO	0.2268	0.2864	**5.9328**	0.3572	0.8999
	RatioDiff	0.2276	**0.2869**	5.8070	**0.3671**	**1.0804**
HPSv2	SDXL	0.2288	0.2849	6.0320	0.3901	0.8750
	SFT	0.2219	0.2824	5.8701	0.3880	0.8194
	Diff.-DPO	**0.2325**	0.2907	6.1544	0.3948	1.1059
	MaPO	0.2295	0.2889	**6.2098**	0.3899	0.9820
	RatioDiff	0.2298	**0.2911**	6.1389	**0.3951**	**1.1210**

5.2 RESULTS

Quantitative Results Table 1reports results for SD1.5. RatioDiff achieves the best scores in 11 of 15 comparisons and improves over the pretrained model on all metrics across all datasets. On Pick-a-Pic v2, it leads HPSv2, Aesthetics, CLIP, and Image Reward, reaching 0.6772 Image Reward, a 6.7% gain over SFT. Similar gains appear on PartiPrompt, while on HPSv2 prompts it leads HPSv2, CLIP, and Image Reward and nearly matches the best Aesthetics score. Table 2 presents the quantitative results for SDXL. the improvements are most evident in reward- and alignment-oriented metrics, where RatioDiff consistently obtains the highest HPSv2, CLIP, and Image Reward scores across all three datasets, accounting for 9 of the 15 best results. In particular, its Image Reward increases to 0.9232 on Pick-a-Pic v2, 2.9% above Diffusion-DPO, and further reaches 1.0804 and 1.1210 on PartiPrompt and HPSv2 prompts. Improvements in PickScore and Aesthetics are comparatively more limited.

Qualitative results. Figure 1 compares SD1.5 generations across four prompts and six methods. While all methods capture the main subject, they differ in prompt-specific details and composition. For the furry-knight prompt, RatioDiff renders a clearer animal-like knight on horseback, whereas some baselines produce more human-like features. For the Chucky–Beetlejuice prompt, the methods share the striped costume but vary in facial appearance, pose, and background. Overall, these examples highlight qualitative differences in prompt adherence and visual composition that complement the aggregate reward scores.

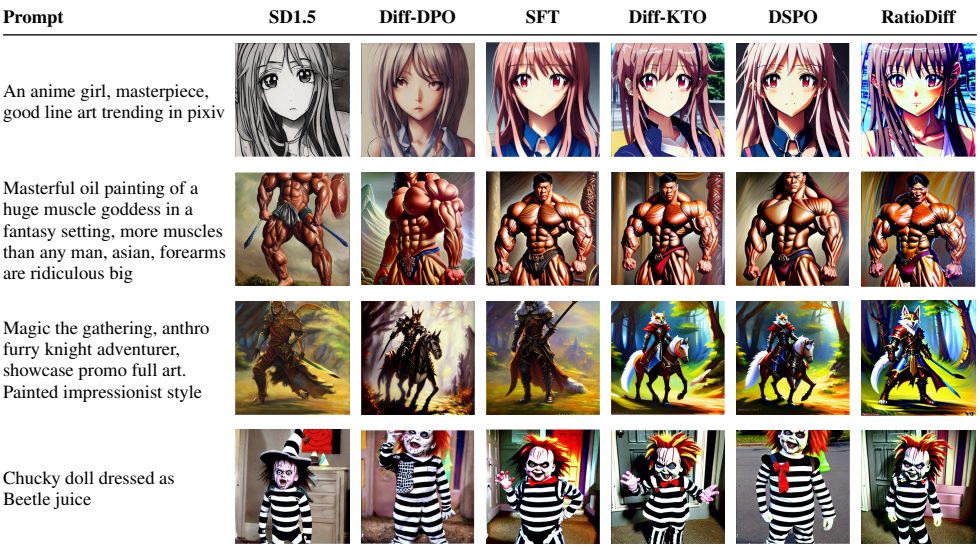


Figure 1: SD1.5 generations for prompts from Pick-a-Pic v2, PartiPrompt, and the HPSv2 prompt set.

Table 3: SD1.5 ablation results on Pick-a-Pic v2 and PartiPrompt. RatioDiff-A: removes the adaptive regularization; RatioDiff-B: removes the adaptive preference scaling.

Dataset	Method	PickScore	HPSv2	Aesthetics	CLIP	Image Reward
Pick-a-Pic v2	RatioDiff-A	0.2114	0.2791	5.6175	0.3427	0.6242
	RatioDiff-B	0.2116	0.2785	5.6314	0.3431	0.6024
	RatioDiff	**0.2124**	**0.2826**	**5.6878**	**0.3458**	**0.6772**
PartiPrompt	RatioDiff-A	0.2168	0.2791	5.5557	0.3376	0.6318
	RatioDiff-B	0.2163	0.2788	5.5638	0.3384	0.6214
	RatioDiff	**0.2178**	**0.2821**	**5.5932**	**0.3402**	**0.6614**

5.3 Ablation Study.

We conduct controlled ablations to isolate the contribution of the main components in RatioDiff. *RatioDiff-A* removes the adaptive regularization, while *RatioDiff-B* removes the adaptive preference scaling $\mathrm{sg}(w_t)s_R$. Table 3 reports the results of these variants on Pick-a-Pic v2 and PartiPrompt. Full RatioDiff consistently outperforms both ablated variants across all five metrics on both datasets, demonstrating that the complete design is important for achieving the strongest overall performance. The performance degradation observed when either adaptive regularization or adaptive preference scaling is removed further confirms that both components contribute meaningfully and complement each other within the full framework.

6 Conclusion

We introduced RatioDiff, an offline preference-learning framework for aligning text-to-image diffusion models through preference ratio matching. RatioDiff derives a reference-corrected transition ratio and optimizes it with a tractable Bregman objective to maintain corrective pressure on strongly misranked pairs. It further combines learned reference-consistency weighting with adaptive regularization to control pair contributions, limit model drift, and preserve preferred-image denoising. These components require no additional denoiser evaluations during training and introduce no extra inference-time overhead. Experiments on SD1.5 and SDXL across three benchmarks demonstrate consistent improvements in preference alignment, text–image consistency, and visual quality.

AI Use Statement

In this work, we used generative AI tools to aid or polish writing, for retrieval and discovery of related work, and to assist in proving mathematical claims. We did not use generative AI tools for

research ideation or execution, drafting sections of the paper, generating synthetic datasets, or any other purposes. We have reviewed and verified all AI-assisted work, including writing revisions, suggested references, and proof steps. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://academic.oup.com/biomet/article-abstract/39/3-4/324/326091.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17400.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.16381.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7441–7451, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hang_Efficient_Diffusion_Training_via_Min-SNR_Weighting_Strategy_ICCV_2023_paper.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4744–4752, 2026. doi: 10.1609/aaai.v40i6.42476. URL https://ojs.aaai.org/index.php/AAAI/article/view/42476. Extended version: https://arxiv.org/abs/2406.06424.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Yeongmin Kim, HeeSun Bae, Byeonghu Na, and Il-Chul Moon. Preference optimization by estimating the ratio of the data distribution. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/559726fdfb19005e368be4ce3d40e3e5-Paper-Conference.pdf.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21696–21707, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b578f2a52a0229873fefc2a4b06377fa-Abstract.html.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. URL https://arxiv.org/abs/2302.12192.

Binxu Li, Minkai Xu, Jiaqi Han, Meihua Dang, and Stefano Ermon. Divergence minimization preference optimization for diffusion model alignment. *arXiv preprint arXiv:2507.07510*, 2025. doi: 10.48550/arXiv.2507.07510. URL https://arxiv.org/abs/2507.07510.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/2c487f8a54cf24c0684c32abc77fed56-Abstract-Conference.html.

Truong Nguyen, Tien-Phat Nguyen, Linh Ngo Van, Duy Minh Ho Nguyen, Khoa D. Doan, and Trung Le. TokenRatio: Principled token-level preference optimization via ratio matching. *CoRR*, abs/2605.12288, 2026. URL https://arxiv.org/abs/2605.12288.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

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. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.01952.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. URL https://arxiv.org/abs/2310.03739.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.18290.

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. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://arxiv.org/abs/2112.10752.

Christoph Schuhmann. LAION-Aesthetics. LAION, 2022. URL https://laion.ai/blog/laion-aesthetics/.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012. doi: 10.1007/s10463-011-0343-8. URL https://www.ism.ac.jp/editsec/aism/pdf/10463_2011_Article_343.pdf.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, abs/2306.09341, 2023. doi: 10.48550/arXiv.2306.09341. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.05977.

Ruojun Xu, Yu Kai, Xuhua Ren, Jiaxiang Cheng, Bing Ma, Tianxiang Zheng, and Qinhlin Lu. Beyond reward margin: Rethinking and resolving likelihood displacement in diffusion models via video generation. *arXiv preprint arXiv:2511.19049*, 2025. URL https://arxiv.org/abs/2511.19049.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. URL https://arxiv.org/abs/2206.10789.

Huaisheng Zhu, Teng Xiao, and Vasant G. Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xyfb9HHvMe.

A Derivations and Training Details

A.1 Proof of Proposition 1

Following TokenRatio (Nguyen et al., 2026), set $y_{t-1} = x_{T-t+1}$ and $y_t = x_{T-t}$. Fix (c, t), with $\beta > 0$, positive reverse-transition densities on common support, and finite normalizers, KL diver-

gences, and expectations. Assume the advantage depends on the history through (c, t, y_{t-1}) and optimize over all transition densities. By Equation 5,

$$\int p_{\mathrm{ref}}(y_t \mid c, y_{t-1}) A_{\mathrm{ref}}([c, y_{<t}], y_t)\, dy_t = 0. \tag{28}$$

Proof of Proposition 1. Equation 6 gives

$$\begin{aligned} p_{\theta^*}(y_t \mid c, y_{t-1}) &= \frac{p_{\mathrm{ref}}(y_t \mid c, y_{t-1}) e^{A_{\mathrm{ref}}([c, y_{<t}], y_t)/\beta}}{Z_t(c, y_{t-1})}, \\ Z_t(c, y_{t-1}) &= \int p_{\mathrm{ref}}(u \mid c, y_{t-1}) e^{A_{\mathrm{ref}}([c, y_{<t}], u)/\beta}\, du. \end{aligned} \tag{29}$$

Indeed, for any transition density p at the same noisy state,

$$\mathbb{E}_p A_{\mathrm{ref}} - \beta D_{\mathrm{KL}}(p \| p_{\mathrm{ref}}) = \beta \log Z_t - \beta D_{\mathrm{KL}}(p \| p_{\theta^*}). \tag{30}$$

Hence

$$\frac{A_{\mathrm{ref}}([c, y_{<t}], y_t)}{\beta} = \log \frac{p_{\theta^*}(y_t \mid c, y_{t-1})}{p_{\mathrm{ref}}(y_t \mid c, y_{t-1})} + \log Z_t(c, y_{t-1}). \tag{31}$$

By Equation 28,

$$\log Z_t(c, y_{t-1}) = D_{\mathrm{KL}}(p_{\mathrm{ref}}(\cdot \mid c, y_{t-1}) \,\|\, p_{\theta^*}(\cdot \mid c, y_{t-1}))\,. \tag{32}$$

By Equation 7,

$$\frac{p_{\mathrm{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\mathrm{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)} = \exp\big(A_{\mathrm{ref}}([c, y_{<t}^w], y_t^w) - A_{\mathrm{ref}}([c, y_{<t}^l], y_t^l)\big)\,. \tag{33}$$

and therefore

$$\log R_{\mathrm{data}}^t = A_{\mathrm{ref}}([c, y_{<t}^l], y_t^l) - A_{\mathrm{ref}}([c, y_{<t}^w], y_t^w). \tag{34}$$

Substituting Equation 31 into the odds gives

$$\frac{p_{\mathrm{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\mathrm{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)} = \left[\frac{p_{\theta^*}(y_t^w \mid c, y_{t-1}^w) p_{\mathrm{ref}}(y_t^l \mid c, y_{t-1}^l)}{p_{\mathrm{ref}}(y_t^w \mid c, y_{t-1}^w) p_{\theta^*}(y_t^l \mid c, y_{t-1}^l)} \frac{Z_t(c, y_{t-1}^w)}{Z_t(c, y_{t-1}^l)}\right]^{\beta}. \tag{35}$$

By Equation 32 and the definition of ω_t following Proposition 1,

$$\omega_t(\theta^*) = \frac{Z_t(c, y_{t-1}^l)}{Z_t(c, y_{t-1}^w)}, \tag{36}$$

and Equation 35 rearranges to

$$\frac{p_{\theta^*}(y_t^w \mid c, y_{t-1}^w)}{p_{\theta^*}(y_t^l \mid c, y_{t-1}^l)} = \frac{p_{\mathrm{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\mathrm{ref}}(y_t^l \mid c, y_{t-1}^l)} \left(\frac{p_{\mathrm{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\mathrm{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)}\right)^{1/\beta} \omega_t(\theta^*). \tag{37}$$

□

A.2 Proof of Proposition 2

Proof of Proposition 2. Let $\widehat{\theta} \in \arg\min_\theta D_h^t(R_{\mathrm{data}}^t, R_\theta^t)$. We have the following decomposition at time step t:

$$D_h^t(R_{\mathrm{data}}^t, R_\theta^t) = \mathbb{E}_{p_{\mathrm{data}}(y_t^w \succ y_t^l \mid c)}\left[B_h(R_{\mathrm{data}}^t, R_\theta^t)\right]. \tag{38}$$

By the non-negativity of the Bregman divergence, $B_h(R_{\mathrm{data}}^t, R_\theta^t)$ attains its minimum value of 0 if and only if $R_{\mathrm{data}}^t = R_\theta^t$ for a given comparison. Assuming that the data distribution has full support, the minimizer $\widehat{\theta}$ therefore satisfies $R_{\mathrm{data}}^t = R_{\widehat{\theta}}^t$ for all covered comparisons at time step t. Substituting the definitions of R_{data}^t and R_θ^t yields the following condition:

$$\frac{p_{\mathrm{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\mathrm{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)} = \left[\frac{p_{\widehat{\theta}}(y_t^l \mid c, y_{t-1}^l)}{p_{\mathrm{ref}}(y_t^l \mid c, y_{t-1}^l)} \frac{p_{\mathrm{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\widehat{\theta}}(y_t^w \mid c, y_{t-1}^w)} \omega_t(\widehat{\theta})\right]^{\beta}. \tag{39}$$

Rearranging, we obtain

$$\frac{p_{\widehat{\theta}}(y_t^w \mid c, y_{t-1}^w)}{p_{\widehat{\theta}}(y_t^l \mid c, y_{t-1}^l)} = \frac{p_{\text{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\text{ref}}(y_t^l \mid c, y_{t-1}^l)} \left(\frac{p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\text{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)} \right)^{1/\beta} \omega_t(\widehat{\theta}). \tag{40}$$

Similarly, Proposition 1 gives the ratio for the optimal policy at the same time step:

$$\frac{p_{\theta^*}(y_t^w \mid c, y_{t-1}^w)}{p_{\theta^*}(y_t^l \mid c, y_{t-1}^l)} = \frac{p_{\text{ref}}(y_t^w \mid c, y_{t-1}^w)}{p_{\text{ref}}(y_t^l \mid c, y_{t-1}^l)} \left(\frac{p_{\text{data}}(y_t^w \succ y_t^l \mid c, y_{<t}^w, y_{<t}^l)}{p_{\text{data}}(y_t^w \prec y_t^l \mid c, y_{<t}^w, y_{<t}^l)} \right)^{1/\beta} \omega_t(\theta^*). \tag{41}$$

Comparing the two expressions reveals that both policies satisfy the same reference-corrected likelihood ratio for every covered comparison at time step t:

$$\frac{p_{\widehat{\theta}}(y_t^w \mid c, y_{t-1}^w)}{p_{\widehat{\theta}}(y_t^l \mid c, y_{t-1}^l)} \frac{1}{\omega_t(\widehat{\theta})} = \frac{p_{\theta^*}(y_t^w \mid c, y_{t-1}^w)}{p_{\theta^*}(y_t^l \mid c, y_{t-1}^l)} \frac{1}{\omega_t(\theta^*)}. \tag{42}$$

Since $\log \omega_t(\theta)$ is the difference between the two reference-to-policy KL terms, the equality above determines the same reference-centered log ratio at each covered noisy-state–action pair. The comparison-support condition and reference centering make that ratio unique; normalization then gives $p_{\widehat{\theta}}(\cdot \mid c, y_{t-1}) = p_{\theta^*}(\cdot \mid c, y_{t-1})$ on the covered support. □

A.3 Proof of Proposition 3: eliminating unknown data ratios

We follow the Bregman expansion and reversed-comparison argument of Theorem 4.4 and Appendix A.4 of TokenRatio (Nguyen et al., 2026). The change of variables is applied to expectations over complete paired states and actions, rather than asserted as a pointwise equality of oppositely ordered comparison probabilities.

Proof of Proposition 3.

By definition:

$$\begin{aligned}
D_h^t(R_{\text{data}}^t, R_\theta^t) &= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h(R_{\text{data}}^t) - h(R_\theta^t) - h'(R_\theta^t)(R_{\text{data}}^t - R_\theta^t) \right] \\
&= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) - h'(R_\theta^t) R_{\text{data}}^t \right] \\
&\quad + \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h(R_{\text{data}}^t) \right] \\
&= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) \right] \\
&\quad - \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'(R_\theta^t) \frac{p_{\text{data}}(y_t^w \prec y_t^l \mid x, y_{<t}^w, y_{<t}^l)}{p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l)} \right] - C \\
&= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) \right] \\
&\quad - \mathbb{E}_{p_{\text{data}}(y_{\le t-1}^w \succ y_{\le t-1}^l \mid x)} \left[h'(R_\theta^t) p_{\text{data}}(y_t^w \prec y_t^l \mid x, y_{<t}^w, y_{<t}^l) \right] - C.
\end{aligned} \tag{43}$$

Next, observe that:

$$\begin{aligned}
& h'(R_\theta^t) p_{\text{data}}(y_t^w \prec y_t^l \mid x, y_{<t}^w, y_{<t}^l) \\
&= p_{\text{data}}(y_t^w \prec y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\big(R_\theta^t(x, y_t^w, y_t^l)\big) \\
&= p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\big(R_\theta^t(x, y_t^l, y_t^w)\big) \\
&= p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\left(\left[\frac{p_\theta(y_t^w \mid [x, y_{<t}^w])\, p_{\text{ref}}(y_t^l \mid [x, y_{<t}^l])}{p_{\text{ref}}(y_t^w \mid [x, y_{<t}^w])\, p_\theta(y_t^l \mid [x, y_{<t}^l])} w_t^{-1} \right]^\beta \right) \\
&= p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\left(\frac{1}{R_\theta^t(x, y_t^w, y_t^l)} \right) \\
&= p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\big((R_\theta^t)^{-1}\big).
\end{aligned} \tag{44}$$

Therefore:

$$\begin{aligned}
& \mathbb{E}_{p_{\text{data}}(y_{\le t-1}^w \succ y_{\le t-1}^l \mid x)} \left[h'(R_\theta^t) p_{\text{data}}(y_t^w \prec y_t^l \mid x, y_{<t}^w, y_{<t}^l) \right] \\
&= \mathbb{E}_{p_{\text{data}}(y_{\le t-1}^w \succ y_{\le t-1}^l \mid x)} \left[p_{\text{data}}(y_t^w \succ y_t^l \mid x, y_{<t}^w, y_{<t}^l) h'\big((R_\theta^t)^{-1}\big) \right] \\
&= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'\big((R_\theta^t)^{-1}\big) \right].
\end{aligned} \tag{45}$$

Substituting the above result into Equation 43 yields:

$$\begin{aligned} D_h^t(R_{\text{data}}^t, R_\theta^t) &= \mathbb{E}_{p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x)} \left[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) - h'\big((R_\theta^t)^{-1}\big) \right] - C \\ &= \mathbb{E}_{p_{\text{data}}(y^w \succ y^l \mid x)} \left[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) - h'\big((R_\theta^t)^{-1}\big) \right] - C. \end{aligned} \tag{46}$$

□

A.4 Gaussian posterior and transition-ratio reduction

For correspondence with the implementation, we use the aliases $\beta_R = \beta$, $\lambda_B = \lambda$, and $s_B = s$ in the derivations below; they are the same parameters as in the main text. For $* \in \{w, l\}$, sample the two branches independently at the same $t \in \mathcal{T}$:

$$\epsilon^* \sim \mathcal{N}(0, I), \qquad x_t^* = \alpha_t x_0^* + \sigma_t \epsilon^*, \qquad x_{t-1}^* \sim q(x_{t-1}^* \mid x_t^*, x_0^*). \tag{47}$$

For adjacent steps, Bayes' rule and the Gaussian forward transitions give

$$\begin{aligned} q(x_{t-1} \mid x_t, x_0) &\propto q(x_t \mid x_{t-1}) q(x_{t-1} \mid x_0) \\ &\propto \exp\left[-\frac{\|x_{t-1} - \alpha_{t-1} x_0\|_2^2}{2\sigma_{t-1}^2} - \frac{\|x_t - \frac{\alpha_t}{\alpha_{t-1}} x_{t-1}\|_2^2}{2\nu_t^2} \right]. \end{aligned} \tag{48}$$

Completing the square and using $x_0 = (x_t - \sigma_t \epsilon)/\alpha_t$ yields

$$\begin{aligned} q(x_{t-1} \mid x_t, x_0) &= \mathcal{N}(\mu_{q,t}, v_t I), \\ \mu_{q,t} &= \frac{\alpha_t \sigma_{t-1}^2}{\alpha_{t-1} \sigma_t^2} x_t + \frac{\alpha_{t-1} \nu_t^2}{\sigma_t^2} x_0 = \frac{\alpha_{t-1}}{\alpha_t} x_t - b_t \epsilon, \\ v_t &= \frac{\nu_t^2 \sigma_{t-1}^2}{\sigma_t^2}, \qquad b_t = \frac{\alpha_{t-1} \nu_t^2}{\alpha_t \sigma_t}. \end{aligned} \tag{49}$$

Replacing x_0 with the denoiser estimate in Equation 17 gives the model and reference transitions, with the same variance v_t: $\epsilon_\theta^* = \epsilon_\theta(x_t^*, t, c)$ and $\epsilon_{\text{ref}}^* = \epsilon_{\text{ref}}(x_t^*, t, c)$.

$$\begin{aligned} \mu_{\theta,t}^* &= \frac{\alpha_{t-1}}{\alpha_t} x_t^* - b_t \epsilon_\theta^*, \\ \mu_{\text{ref},t}^* &= \frac{\alpha_{t-1}}{\alpha_t} x_t^* - b_t \epsilon_{\text{ref}}^*, \\ p_\theta(x_{t-1}^* \mid x_t^*, c) &= \mathcal{N}(\mu_{\theta,t}^*, v_t I), \\ p_{\text{ref}}(x_{t-1}^* \mid x_t^*, c) &= \mathcal{N}(\mu_{\text{ref},t}^*, v_t I). \end{aligned} \tag{50}$$

At one branch, the common Gaussian normalization cancels:

$$\log \frac{p_\theta(x_{t-1} \mid x_t, c)}{p_{\text{ref}}(x_{t-1} \mid x_t, c)} = \frac{\|x_{t-1} - \mu_{\text{ref},t}\|_2^2 - \|x_{t-1} - \mu_{\theta,t}\|_2^2}{2 v_t}. \tag{51}$$

The KL correction in ω_t cancels its quadratic term:

$$\begin{aligned} &\log \frac{p_\theta(x_{t-1} \mid x_t, c)}{p_{\text{ref}}(x_{t-1} \mid x_t, c)} + D_{\text{KL}}(p_{\text{ref}} \| p_\theta) \\ &= \frac{\|x_{t-1} - \mu_{\text{ref},t}\|_2^2 - \|x_{t-1} - \mu_{\theta,t}\|_2^2 + \|\mu_{\theta,t} - \mu_{\text{ref},t}\|_2^2}{2 v_t} \\ &= \frac{\langle \mu_{\theta,t} - \mu_{\text{ref},t}, x_{t-1} - \mu_{\text{ref},t} \rangle}{v_t}. \end{aligned} \tag{52}$$

Here the KL is between transitions at the fixed state (x_t, c). For $R_{\theta,t}^{\text{tr}} = R_\theta^{T-t+1}$, Equation 10 reads

$$\log R_{\theta,t}^{\text{tr}} = \beta_R \left[\log \frac{p_\theta(x_{t-1}^l \mid x_t^l, c)}{p_{\text{ref}}(x_{t-1}^l \mid x_t^l, c)} - \log \frac{p_\theta(x_{t-1}^w \mid x_t^w, c)}{p_{\text{ref}}(x_{t-1}^w \mid x_t^w, c)} + \log \omega_t \right], \tag{53}$$

where $\log \omega_t$ is the rejected-minus-preferred KL difference in Equation 10. By Equation 52, this becomes

$$\log R^{\mathrm{tr}}_{\theta,t} = \frac{\beta_R}{v_t}\Big[\langle \mu^l_{\theta,t} - \mu^l_{\mathrm{ref},t}, x^l_{t-1} - \mu^l_{\mathrm{ref},t}\rangle - \langle \mu^w_{\theta,t} - \mu^w_{\mathrm{ref},t}, x^w_{t-1} - \mu^w_{\mathrm{ref},t}\rangle\Big]. \tag{54}$$

Write $q^*_t = q(x^*_{t-1} \mid x^*_t, x^*_0)$, $e^*_t = \epsilon^* - \epsilon^*_{\mathrm{ref}}$, and $\delta^*_t = \epsilon^*_\theta - \epsilon^*_{\mathrm{ref}}$. Since $\mu^*_{q,t} - \mu^*_{\mathrm{ref},t} = -b_t e^*_t$ and $\mu^*_{\theta,t} - \mu^*_{\mathrm{ref},t} = -b_t \delta^*_t$,

$$\begin{aligned}\mathbb{E}_{q^*_t}\left[\log \frac{p_\theta}{p_{\mathrm{ref}}} + D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_\theta)\right] &= \frac{\langle -b_t\delta^*_t, -b_t e^*_t\rangle}{v_t} \\ &= 2\chi_t \langle e^*_t, \delta^*_t\rangle, \qquad \chi_t = \frac{b_t^2}{2v_t}.\end{aligned} \tag{55}$$

Thus, averaging the two sampled transitions gives

$$\mathbb{E}_{q^w_t q^l_t} \log R^{\mathrm{tr}}_{\theta,t} = 2\beta_R \chi_t \left[\langle e^l_t, \delta^l_t\rangle - \langle e^w_t, \delta^w_t\rangle\right] = \beta_R m_{\theta,t}. \tag{56}$$

The practical ratio is the exponential of this posterior mean:

$$\begin{aligned} m_{\theta,t} &= 2\chi_t \left[\langle e^l_t, \delta^l_t\rangle - \langle e^w_t, \delta^w_t\rangle\right], \\ R_{\theta,t} &= \exp(\beta_R m_{\theta,t}). \end{aligned} \tag{57}$$

For completeness, let $E^*_{\theta,t} = \|\epsilon^* - \epsilon^*_\theta\|_2^2$. The same Gaussian calculation also gives, on either branch,

$$\begin{aligned} \mathbb{E}_{q^*_t} \log \frac{p_\theta}{p_{\mathrm{ref}}} &= \chi_t (E^*_{\mathrm{ref},t} - E^*_{\theta,t}), \\ D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_\theta) = \chi_t \|\delta^*_t\|_2^2, &\qquad D_{\mathrm{KL}}(q^*_t \| p_\theta) = \chi_t E^*_{\theta,t}. \end{aligned} \tag{58}$$

The first identity follows from $\mathbb{E}_{q^*_t}\|x^*_{t-1} - \mu\|_2^2 = \|\mu^*_{q,t} - \mu\|_2^2 + d v_t$.

The two posterior draws are independent. Equation 52 therefore gives, conditional on the clean pair, timestep, and sampled noises,

$$\log R^{\mathrm{tr}}_{\theta,t} \sim \mathcal{N}(\beta_R m_{\theta,t}, V_{\theta,t}), \qquad V_{\theta,t} = 2\beta_R^2 \chi_t \big(\|\delta^w_t\|_2^2 + \|\delta^l_t\|_2^2\big). \tag{59}$$

Hence $R_{\theta,t} = e^{\mathbb{E}\log R^{\mathrm{tr}}_{\theta,t}}$, while $\mathbb{E}R^{\mathrm{tr}}_{\theta,t} = R_{\theta,t} e^{V_{\theta,t}/2}$. Using $R_{\theta,t}$ inside ℓ_h therefore defines a Gaussian surrogate for the exact transition loss. These identities require $v_t > 0$; the deterministic endpoints are excluded by $\mathcal{T}$.

A.5 Basu gradients and loss scaling

The per-comparison loss in Equation 13 is $\ell_h(R) = h'(R)R - h(R) - h'(R^{-1})$. For the Basu generator in Equation 14,

$$\begin{aligned} h'(R) &= \frac{(1+\lambda)R^\lambda - 1}{s_B \lambda(1+\lambda)}, \\ h''(R) &= \frac{R^{\lambda-1}}{s_B} > 0. \end{aligned} \tag{60}$$

Strict convexity makes h a valid Bregman generator. Differentiating with respect to the log-ratio gives

$$\begin{aligned} \frac{d\ell_h(R)}{d\log R} &= R^2 h''(R) + R^{-1} h''(R^{-1}) \\ &= \frac{R^{1+\lambda} + R^{-\lambda}}{s_B} > 0. \end{aligned} \tag{61}$$

Because $R_{\theta,t}$ is the rejected-to-preferred ratio, minimizing this loss pushes it downward. The shape parameter λ changes the response away from $R = 1$, including both tails. With $R_{\theta,t} = e^{\beta m_{\theta,t}}$, the unscaled slope at $m_{\theta,t} = 0$ is $2\beta/s_B$. We therefore use the multiplier from Section 4.3, $s_R = s_B S_R/(2\beta)$, so that

$$\begin{aligned} \frac{\partial}{\partial m_{\theta,t}}\big[s_R \ell_h(R_{\theta,t})\big] &= \frac{\beta s_R}{s_B}\left(R^{1+\lambda}_{\theta,t} + R^{-\lambda}_{\theta,t}\right) \\ &= \frac{S_R}{2}\left(R^{1+\lambda}_{\theta,t} + R^{-\lambda}_{\theta,t}\right). \end{aligned} \tag{62}$$

At reference initialization, $R_{\theta,t} = 1$, and hence

$$\begin{aligned} \frac{\partial}{\partial m_{\theta,t}}\big[s_R \ell_h(R_{\theta,t})\big]\Big|_{m_{\theta,t}=0} &= S_R, \\ \frac{\partial}{\partial m_{\theta,t}}\big[\mathrm{sg}(w_t) s_R \ell_h(R_{\theta,t})\big]\Big|_{m_{\theta,t}=0} &= \mathrm{sg}(w_t) S_R. \end{aligned} \tag{63}$$

Thus β controls the ratio's nonlinear sensitivity, while S_R sets its initial update strength; s_B is canceled by s_R. Indeed, the entire calibrated loss is independent of s_B:

$$s_R \ell_h(R) = \frac{S_R}{2\beta}\left[\frac{R^{1+\lambda}-1}{1+\lambda} - \frac{R^{-\lambda}-1}{\lambda}\right]. \tag{64}$$

The positive slope in Equation 61 retains a corrective signal for misranked pairs; $\mathrm{sg}(w_t)$ adjusts each pair's contribution without changing this calibration.

The continuous $\lambda \to 0$ limit used in the implementation follows from $(R^\lambda - 1)/\lambda \to \log R$:

$$\begin{aligned} h(R)|_{\lambda=0} &= \frac{R \log R}{s_B}, \qquad \ell_h(R)|_{\lambda=0} = \frac{R - 1 + \log R}{s_B}, \\ \frac{\partial}{\partial m_{\theta,t}}\big[s_R \ell_h(R_{\theta,t})\big]\Big|_{\lambda=0} &= \frac{S_R}{2}(R_{\theta,t} + 1). \end{aligned} \tag{65}$$

This limit keeps the initial slope S_R.

A.6 Reference-Consistency Confidence and Pair Weighting

At a sampled timestep t, the head in Equation 19 receives only the noisy pair, frozen-reference predictions, and timestep. For $* \in \{w, l\}$, let

$$\begin{aligned} F_t^* &= [x_t^*; \epsilon_{\mathrm{ref}}^*], \\ u_t^* &= [\mu_{\mathrm{sp}}(F_t^*); \sigma_{\mathrm{sp}}(F_t^*)], \\ Z_t &= [u_t^w; u_t^l; u_t^w - u_t^l; |u_t^w - u_t^l|; \tau(t)]. \end{aligned} \tag{66}$$

Here, μ_{sp} and σ_{sp} pool over spatial coordinates, and $\tau(t)$ is a sinusoidal timestep embedding. The signed and absolute differences encode branch order and gap size. With layer normalization and two SiLU hidden layers, the predictor is

$$\begin{aligned} h_{1,t} &= \mathrm{SiLU}(W_1\,\mathrm{LN}(\mathrm{sg}(Z_t)) + b_1), \\ h_{2,t} &= \mathrm{SiLU}(W_2 h_{1,t} + b_2), \\ p_{\phi,t} &= \mathrm{sigm}(W_3 h_{2,t} + b_3). \end{aligned} \tag{67}$$

Thus g_ϕ is the network before the final sigmoid. The inputs are detached from the denoiser; ϕ contains the affine and normalization parameters.

Human labels compare clean images and do not label individual noisy steps. Using the reference errors in Equation 20, the main-text proxy target is

$$d_t = \begin{cases} 1, & E_{\mathrm{ref},t}^w < E_{\mathrm{ref},t}^l, \\ 0, & E_{\mathrm{ref},t}^w > E_{\mathrm{ref},t}^l, \\ \frac{1}{2}, & E_{\mathrm{ref},t}^w = E_{\mathrm{ref},t}^l. \end{cases} \qquad \widetilde{d}_t = (1-\eta)d_t + \frac{\eta}{2}, \quad 0 \le \eta < 1. \tag{68}$$

The head minimizes the same binary cross-entropy as Equation 21, per sampled pair:

$$\begin{aligned} \ell_{\mathrm{conf},t} &= -\widetilde{d}_t \log p_{\phi,t} - (1 - \widetilde{d}_t)\log(1 - p_{\phi,t}), \\ \mathcal{L}_{\mathrm{conf}} &= \mathbb{E}[\ell_{\mathrm{conf},t}]. \end{aligned} \tag{69}$$

For fixed $Z_t = Z$, the expected target is

$$\begin{aligned} \mathbb{E}[\widetilde{d}_t \mid Z_t = Z] = (1-\eta)\Big[&\Pr(E_{\mathrm{ref},t}^w < E_{\mathrm{ref},t}^l \mid Z_t = Z) \\ &+ \tfrac{1}{2}\Pr(E_{\mathrm{ref},t}^w = E_{\mathrm{ref},t}^l \mid Z_t = Z)\Big] + \frac{\eta}{2}. \end{aligned} \tag{70}$$

For a predicted probability $p \in (0,1)$, differentiating the conditional cross-entropy gives

$$\frac{\partial}{\partial p}\mathbb{E}[\ell_{\mathrm{conf},t} \mid Z_t = Z] = \frac{p - \mathbb{E}[\widetilde{d}_t \mid Z_t = Z]}{p(1-p)}. \tag{71}$$

Thus the population optimum is $p = \mathbb{E}[\widetilde{d}_t \mid Z_t = Z]$ when it lies in $(0,1)$; boundary optima are approached by the sigmoid. This estimates smoothed reference consistency, not a human preference probability at timestep t. A finite model need not reach the population optimum.

The floor and batch normalization in Equation 22 give

$$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} w_i = 1, \qquad w_{\min} \le w_i \le \frac{1}{w_{\min}}, \qquad \frac{w_i}{w_k} = \frac{\widetilde{w}_i}{\widetilde{w}_k}. \tag{72}$$

The floor retains every pair; normalization preserves relative weights and fixes their batch mean. With one pair, $w_i = 1$. For $n = |\mathcal{B}|$, detaching w_i in the preference loss gives

$$\begin{aligned} \widehat{\mathcal{L}}_{\mathrm{pref}} &= \frac{1}{n}\sum_i \mathrm{sg}(w_i) s_R \ell_h(R_{\theta,i}), \\ \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{pref}} &= \frac{S_R}{2n}\sum_i \mathrm{sg}(w_i)\big(R_{\theta,i}^{1+\lambda} + R_{\theta,i}^{-\lambda}\big)\nabla_\theta m_{\theta,i}, \\ \nabla_\phi \widehat{\mathcal{L}}_{\mathrm{pref}} &= 0. \end{aligned} \tag{73}$$

The head therefore learns from $\mathcal{L}_{\mathrm{conf}}$ alone; preference optimization cannot lower its own loss by changing the weights. Since w_i depends on the pair and its batch, the exact unweighted ratio guarantee of Proposition 2 does not automatically extend to this weighted surrogate.

A.7 Geometry of Reference and Winner Regularization

The two terms in Equations 24 and 25 constrain different aspects of denoising. For either branch, the shared Gaussian covariance and Equation 50 give

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{ref}}(\cdot \mid x_t^*, c) \,\|\, p_\theta(\cdot \mid x_t^*, c)) &= \frac{\|\mu_{\theta,t}^* - \mu_{\mathrm{ref},t}^*\|_2^2}{2v_t} \\ &= \frac{b_t^2}{2v_t}\|\delta_t^*\|_2^2 = \chi_t \|\delta_t^*\|_2^2. \end{aligned} \tag{74}$$

Thus $\mathcal{L}_{\mathrm{ref},t}$ is the sum of the two transition KL terms. In contrast, Equation 57 uses only the projection of each update δ_t^* onto the reference residual e_t^*. Cauchy–Schwarz makes the distinction explicit:

$$\begin{aligned} m_{\theta,t}^2 &= 4\chi_t^2\big(-\langle e_t^w, \delta_t^w\rangle + \langle e_t^l, \delta_t^l\rangle\big)^2 \\ &\le 4\chi_t^2\big(\|e_t^w\|_2^2 + \|e_t^l\|_2^2\big)\big(\|\delta_t^w\|_2^2 + \|\delta_t^l\|_2^2\big) \\ &= 4\chi_t\big(\|e_t^w\|_2^2 + \|e_t^l\|_2^2\big)\mathcal{L}_{\mathrm{ref},t}. \end{aligned} \tag{75}$$

For $\chi_t > 0$ and nonzero reference residuals, this is equivalent to $\mathcal{L}_{\mathrm{ref},t} \ge m_{\theta,t}^2/[4\chi_t(\|e_t^w\|_2^2 + \|e_t^l\|_2^2)]$. Equality holds when (δ_t^w, δ_t^l) is proportional to $(-e_t^w, e_t^l)$; components orthogonal to this direction change the reference KL but not the ratio margin. When both residuals vanish, $m_{\theta,t} = 0$ while $\mathcal{L}_{\mathrm{ref},t}$ can still penalize drift. These KL terms constrain transitions at sampled states.

A bounded correction for the preferred target. The winner term acts directly on the preferred noise prediction. To set its correction, define the denoising errors and preferred-minus-rejected gaps:

$$\begin{aligned} E_{\theta,t}^* &= \|\epsilon^* - \epsilon_\theta^*\|_2^2, \qquad * \in \{w, l\}, \\ \Delta_{\theta,t} &= E_{\theta,t}^w - E_{\theta,t}^l, \\ \Delta_{\mathrm{ref},t} &= E_{\mathrm{ref},t}^w - E_{\mathrm{ref},t}^l. \end{aligned} \tag{76}$$

The reference errors are defined in Equation 20. Only the combination $\beta_D/(2T_D)$ enters the score, so we combine it into one positive sensitivity parameter $\gamma_D = \beta_D/(2T_D) > 0$. The gate is

$$\begin{aligned} p_{D,t} &= \mathrm{sigm}[-\gamma_D(\Delta_{\theta,t} - \Delta_{\mathrm{ref},t})], \\ \kappa_t &= \kappa_{\max}(1 - \mathrm{sg}(p_{D,t})), \qquad 0 \le \kappa_{\max} < 1. \end{aligned} \tag{77}$$

As $\Delta_{\theta,t} - \Delta_{\mathrm{ref},t}$ decreases, the preferred-versus-rejected denoising gap improves relative to the reference, $p_{D,t}$ increases, and κ_t decreases. The score measures this denoising-gap improvement; it is distinct from the reference-corrected ratio margin $m_{\theta,t}$. Detachment lets both branches set the gate without receiving gradients through it. Since $0 \le \kappa_t \le \kappa_{\max}$ and $\kappa_{\max} < 1$, Equation 25 can be rearranged as

$$\begin{aligned} r_{\mathrm{win},t} &= (1-\kappa_t)\delta_t^w - e_t^w = (1-\kappa_t)(\epsilon_\theta^w - \epsilon_{\mathrm{tar},t}^w), \\ \epsilon_{\mathrm{tar},t}^w &= \frac{\epsilon^w - \kappa_t \epsilon_{\mathrm{ref}}^w}{1-\kappa_t} = \epsilon_{\mathrm{ref}}^w + \frac{e_t^w}{1-\kappa_t}, \\ \mathcal{L}_{\mathrm{win},t} &= (1-\kappa_t)^2 \|\epsilon_\theta^w - \epsilon_{\mathrm{tar},t}^w\|_2^2. \end{aligned} \tag{78}$$

If $e_t^w \neq 0$, increasing κ_t moves the target beyond ϵ^w along e_t^w and reduces the quadratic prefactor. The amplification factor $1/(1-\kappa_t)$ is bounded by $1/(1-\kappa_{\max})$. Holding the gate fixed during differentiation gives

$$\begin{aligned} \nabla_{\epsilon_\theta^w} \mathcal{L}_{\mathrm{win},t} &= 2(1-\kappa_t) r_{\mathrm{win},t}, \\ \nabla_{\epsilon_\theta^l} \mathcal{L}_{\mathrm{win},t} &= 0, \\ \nabla^2_{\epsilon_\theta^w} \mathcal{L}_{\mathrm{win},t} &= 2(1-\kappa_t)^2 I \succeq 2(1-\kappa_{\max})^2 I. \end{aligned} \tag{79}$$

The frozen-gate loss has a unique output-space minimum at $\epsilon_{\mathrm{tar},t}^w$. The rejected branch sets the gate but receives no direct output gradient from this term. The Hessian bound applies to the current detached computation; the gate is recomputed after a model update. When $p_{D,t} \to 1$ or $\kappa_{\max} = 0$, $\kappa_t \to 0$ and the term becomes ordinary preferred-image denoising. At reference initialization, $p_{D,t} = 1/2$ and $\kappa_t = \kappa_{\max}/2$: the loss value is $\|e_t^w\|_2^2$, while its preferred-output gradient is $-2(1-\kappa_{\max}/2)e_t^w$.

A.8 Output Gradients and Auxiliary Budget Bounds

All three losses act on the same winner and loser denoiser outputs. Comparing their gradients with respect to $(\epsilon_\theta^w, \epsilon_\theta^l)$ therefore puts them in the same coordinates, before gradients pass through the shared network. For a fixed corrupted pair, hold the forward noises, reference predictions, w_t, and κ_t fixed. Write $g_{k,t} = \nabla_{(\epsilon_\theta^w, \epsilon_\theta^l)} \mathcal{L}_{k,t}$ and $G_{k,t} = \|g_{k,t}\|_2$ for $k \in \{\mathrm{pref}, \mathrm{ref}, \mathrm{win}\}$. We use the preference norm as the scale for both regularizers. For $k \in \{\mathrm{ref}, \mathrm{win}\}$, set

$$a_{k,t} = \mathrm{sg}\left[\min\left(a_{\max}, \frac{\rho_k G_{\mathrm{pref},t}}{\max(G_{k,t}, \varepsilon_g)}\right)\right], \qquad \rho_k \ge 0, \quad a_{\max}, \varepsilon_g > 0. \tag{80}$$

Here ρ_k sets the allowed fraction of the preference-gradient norm, $a_{\max}$ caps the coefficient, and ε_g prevents division by zero. Detaching $a_{k,t}$ makes the scaled regularizer's gradient exactly $a_{k,t} g_{k,t}$. Thus the rule limits the regularizer's actual gradient contribution in denoiser-output space. The norms below have analytic expressions that reuse the existing predictions, so this adaptation requires no additional denoiser evaluations.

Analytic output gradients. For the per-pair preference term $\mathcal{L}_{\mathrm{pref},t} = \mathrm{sg}(w_t) s_R \ell_h(R_{\theta,t})$, Equations 57 and 62 give

$$\begin{aligned} \nabla_{\epsilon_\theta^w} m_{\theta,t} &= -2\chi_t e_t^w, \\ \nabla_{\epsilon_\theta^l} m_{\theta,t} &= 2\chi_t e_t^l, \\ \frac{\partial \mathcal{L}_{\mathrm{pref},t}}{\partial m_{\theta,t}} &= \mathrm{sg}(w_t) \frac{S_R}{2} \left(R_{\theta,t}^{1+\lambda} + R_{\theta,t}^{-\lambda}\right). \end{aligned} \tag{81}$$

Their product gives the positive preference-gradient scale

$$C_t = \mathrm{sg}(w_t) S_R \chi_t \left(R_{\theta,t}^{1+\lambda} + R_{\theta,t}^{-\lambda}\right) > 0. \tag{82}$$

Differentiating all three terms gives

$$\begin{aligned} g_{\mathrm{pref},t} &= C_t(-e_t^w, e_t^l), \\ G_{\mathrm{pref},t} &= C_t\sqrt{\|e_t^w\|_2^2 + \|e_t^l\|_2^2}, \\ g_{\mathrm{ref},t} &= 2\chi_t(\delta_t^w, \delta_t^l), \\ G_{\mathrm{ref},t} &= 2\chi_t\sqrt{\|\delta_t^w\|_2^2 + \|\delta_t^l\|_2^2}, \\ g_{\mathrm{win},t} &= (2(1-\kappa_t)r_{\mathrm{win},t}, 0), \\ G_{\mathrm{win},t} &= 2(1-\kappa_t)\|r_{\mathrm{win},t}\|_2. \end{aligned} \tag{83}$$

The winner gradient uses the frozen-gate residual in Equation 78. These are per-pair output-gradient norms before batch averaging; they reuse denoiser outputs already computed for the losses. In particular, a negative preference-gradient step moves the winner along $+e_t^w$ and the loser along $-e_t^l$ in these output coordinates. The reference gradient pulls both outputs toward the frozen model, while the winner gradient acts only on the preferred branch. The controller in Equation 80 uses the norms $G_{k,t}$ to limit each auxiliary relative to the preference signal.

Proposition 4 (Per-pair auxiliary gradient bound)**.** *For the detached coefficients in Equation 80, each auxiliary contributes at most a fraction ρ_k of the local preference-gradient norm: $\|a_{k,t}g_{k,t}\|_2 \le \rho_k G_{\mathrm{pref},t}$. If $\rho = \rho_{\mathrm{ref}} + \rho_{\mathrm{win}} < 1$ and $G_{\mathrm{pref},t} > 0$, then $-g_{\mathrm{total},t}$, where $g_{\mathrm{total},t} = g_{\mathrm{pref},t} + a_{\mathrm{ref},t}g_{\mathrm{ref},t} + a_{\mathrm{win},t}g_{\mathrm{win},t}$, is a first-order descent direction for the local preference loss in output space.*

Proof. The denominator floor ε_g avoids division by zero, and the cap $a_{\max}$ can only reduce $a_{k,t}$. Thus, for $k \in \{\mathrm{ref}, \mathrm{win}\}$,

$$\begin{aligned} \|a_{k,t}g_{k,t}\|_2 &= a_{k,t}G_{k,t} \\ &\le \frac{\rho_k G_{\mathrm{pref},t}G_{k,t}}{\max(G_{k,t}, \varepsilon_g)} \le \rho_k G_{\mathrm{pref},t}. \end{aligned} \tag{84}$$

The last step uses $G_{k,t}/\max(G_{k,t}, \varepsilon_g) \le 1$ and also holds when $G_{k,t} = 0$. If $G_{\mathrm{pref},t} = 0$, both auxiliary coefficients vanish. Otherwise, the triangle inequality shows that the two auxiliaries together change the preference gradient by at most $\rho G_{\mathrm{pref},t}$:

$$\|g_{\mathrm{total},t} - g_{\mathrm{pref},t}\|_2 \le \sum_{k\in\{\mathrm{ref},\mathrm{win}\}} a_{k,t}G_{k,t} \le \rho G_{\mathrm{pref},t}.$$

Cauchy–Schwarz then gives

$$\begin{aligned} \langle g_{\mathrm{pref},t}, g_{\mathrm{total},t}\rangle &\ge G_{\mathrm{pref},t}^2 - G_{\mathrm{pref},t}\|g_{\mathrm{total},t} - g_{\mathrm{pref},t}\|_2 \\ &\ge (1-\rho)G_{\mathrm{pref},t}^2 > 0. \end{aligned} \tag{85}$$

Because $\rho < 1$, the combined gradient still points partly along $g_{\mathrm{pref},t}$. A sufficiently small step along $-g_{\mathrm{total},t}$ therefore decreases the preference loss for this pair in output space. □

The bound concerns one pair in denoiser-output space. Detaching each $a_{k,t}$ makes the auxiliary gradient exactly $a_{k,t}g_{k,t}$; without detachment, differentiating the coefficient would add $\mathcal{L}_{k,t}\nabla_{(\epsilon_\theta^w, \epsilon_\theta^l)}a_{k,t}$. Backpropagation then maps each output gradient to $J_t^\top g_{k,t}$, where $J_t = \partial(\epsilon_\theta^w, \epsilon_\theta^l)/\partial\theta$.

B Training Procedure

Sections 4.3 and 4.4 require the current and frozen-reference predictions on the same corrupted pair. The head and controllers reuse these outputs; they do not require a reward model, a critic, or additional reverse-action samples. With batch size $n = |\mathcal{B}|$, one training step proceeds as follows.

1. Sample n ordered image pairs. For each pair i, draw $t_i \sim \mathcal{U}(\mathcal{T})$ and independent standard Gaussian noises $\epsilon_i^w, \epsilon_i^l$, sharing t_i between its two branches. Construct $x_{t_i,i}^* = \alpha_{t_i}x_{0,i}^* + \sigma_{t_i}\epsilon_i^*$ for $* \in \{w, l\}$.

2. Evaluate $\epsilon^*_{\theta,i}$ and the frozen $\epsilon^*_{\mathrm{ref},i}$ on these inputs. Form e^*_i, δ^*_i, and the denoising errors, and evaluate $z_i = \log R_{\theta,i} = \beta m_{\theta,i}$ using the inner-product expression in Equation 57.
3. Construct the detached statistics Z_i, reference-ordering target $\widetilde{d}_i$, confidence loss $\ell_{\mathrm{conf},i}$, and normalized weight w_i from Equations 66, 68, 69, and 22. Compute $\mathcal{L}_{\mathrm{pref},i} = \mathrm{sg}(w_i) s_R \ell_h(e^{z_i})$, where $s_R = s_B S_R/(2\beta)$.
4. Compute $\mathcal{L}_{\mathrm{ref},i}$ and the detached winner gate κ_i, then $\mathcal{L}_{\mathrm{win},i}$ using Equations 24, 77, and 25. Evaluate the three output-gradient norms in Equation 83 and detach the resulting $a_{\mathrm{ref},i}$ and $a_{\mathrm{win},i}$ from Equation 80.
5. Evaluate Equation 27 as a batch mean and backpropagate. The denoiser receives the preference and two auxiliary gradients; the confidence head receives only $\lambda_{\mathrm{conf}} n^{-1} \sum_{i\in\mathcal{B}} \nabla_\phi \ell_{\mathrm{conf},i}$. Apply the chosen optimizer to θ and ϕ, keeping the reference model and image/text encoders frozen.

The parameters β, s_B, λ, S_R, $w_{\min}$, η, λ_{conf}, $\kappa_{\max}$, γ_D, ρ_{ref}, ρ_{win}, $a_{\max}$, and ε_g are fixed hyperparameters for a run, not outputs of the confidence head. The calibrated slope S_R sets the initial derivative with respect to the margin $m_{\theta,i}$, independently of s_B and β. All reference features and targets are independent of the trainable denoiser.

Compute ratios through $z_i = \log R_{\theta,i}$, evaluating powers as $R_{\theta,i}^{1+\lambda} = e^{(1+\lambda)z_i}$ and $R_{\theta,i}^{-\lambda} = e^{-\lambda z_i}$. The derivations assume the unclipped loss and consistent coordinate reductions. If numerical guards modify these expressions, the controller must use the corresponding modified derivatives for its gradient budget to remain valid. The head, reference model, and controllers are used only during training.

C Experimental and Implementation Details

C.1 Evaluation Protocol

Datasets. Both Stable Diffusion v1.5 (SD1.5) (Rombach et al., 2022) and Stable Diffusion XL 1.0 (SDXL) (Podell et al., 2024) are fine-tuned on 851,293 non-tied Pick-a-Pic v2 image pairs (Kirstain et al., 2023). Pick-a-Pic contains prompts entered by real users and their preferences between two generated images. For evaluation, we use 500 distinct prompts from its `test_unique` split. PartiPrompt (Yu et al., 2022) contributes 1,632 curated English prompts covering varied subjects, attributes, compositions, and levels of difficulty. The third set contains 400 prompts from the HPD v2 test split (Wu et al., 2023), denoted *HPSv2* in our tables; HPD v2 is a separate collection of generated images with human preference annotations. Thus, Pick-a-Pic evaluates prompts from the training dataset's source, while PartiPrompt and HPD v2 provide two additional prompt distributions. Only the prompt text is used to generate evaluation images.

Baselines. The pretrained SD1.5 and SDXL models show performance before preference fine-tuning. Supervised fine-tuning (SFT) trains on the preferred images alone, whereas Diffusion-DPO (Wallace et al., 2024) uses both images in each pair and a reference-relative logistic preference loss. For SD1.5, we additionally compare with Diffusion-KTO (Li et al., 2024), which learns from desirable and undesirable image labels; DMPO (Li et al., 2025), which minimizes a reverse-KL-based preference objective; and DSPO (Zhu et al., 2025), which uses preference-aware score matching. For SDXL, we include MaPO (Hong et al., 2026), which combines a preferred-versus-rejected likelihood margin with a preferred-image likelihood term without a frozen reference model. We use released checkpoints or official implementations when available, retrain DSPO with its official code, and reimplement the SD1.5 DMPO baseline as detailed in Appendix C.4. All comparisons are made within the same backbone.

Metrics. PickScore (Kirstain et al., 2023) is a CLIP-based scorer trained on Pick-a-Pic user comparisons; it measures how strongly an image matches the preferences associated with its prompt. HPSv2 (Wu et al., 2023) is trained on HPD v2 human comparisons and provides a second preference-oriented image–text score. ImageReward (Xu et al., 2023) is a separately trained reward model that assigns a scalar human-preference score conditioned on the prompt. These three scorers assess

preference from different training data and models. CLIP Score (Radford et al., 2021) measures semantic image–text similarity, whereas LAION Aesthetics Score (Schuhmann, 2022) predicts visual appeal from the image alone. The latter two help distinguish prompt adherence and visual quality from preference-score improvements.

Scoring protocol. We generate one image per prompt and method, using the same prompts and generation settings within each backbone. SD1.5 uses 512×512 resolution, 50 sampling steps, classifier-free guidance of 7.5, and seed $42 + i$ for prompt index i. For every metric, each entry in Tables 1 and 2 is the arithmetic mean of its scores across all prompts in that set; higher is better. The Aesthetics scorer uses only the generated image, while the other four also receive its prompt.

C.2 Implementation Details

We implement RatioDiff using the Diffusion-DPO training framework (Wallace et al., 2024) and train SD1.5 on four NVIDIA A100 80 GB GPUs with Hugging Face Accelerate. We fully fine-tune the UNet while keeping the reference model, VAE, and text encoder frozen. Training uses AdamW with an effective batch size of 64 pairs, accumulating gradients over 16 steps with one pair per GPU. The learning rates are 2.5×10^{-6} for the UNet and 10^{-4} for the confidence head, with 200 steps of linear warmup followed by a constant schedule. We train for 13,302 steps on 851,293 preference pairs at 512×512 resolution, using bfloat16 mixed precision, gradient checkpointing, gradient clipping at 1.0, and seed 42.

The confidence head is a two-layer MLP with 128 hidden units per layer and SiLU activations. We use label smoothing of 0.05, a minimum raw pair weight of 0.05, and normalize detached weights across each global microbatch. The reference and winner gradient budgets are 0.10 and 0.25, with winner-auxiliary parameters $\gamma_D = 0.005$ and $\kappa_{\max} = 0.25$. The ratio objective uses $\beta = 500$, target initial margin slope $S_R = 2{,}500$, the continuous $\lambda = 0$ Basu limit, and $s_B = 4$; thus $s_R = 10$ under mean reduction over latent coordinates. Our implementation samples posterior actions with exact Gaussian state correction, shares timesteps and forward noise within pairs, and computes auxiliary coefficients from gradient norms aggregated across GPUs. These choices differ from the posterior-averaged, per-pair-budget formulation.

C.3 SDXL Implementation

Diffusion-DPO suggests that, for a strong pretrained model such as SDXL, the offline preference data may not always provide samples of higher quality than those already produced by the original model. In such cases, forcing the model toward the training images can weaken its pretrained generation capability. This observation motivates using the pretrained model as a reference, providing a stronger and more reliable target for preserving generation quality during preference optimization. The same consideration also underlies the SDXL-specific treatment adopted in DSPO (Zhu et al., 2025). For RatioDiff, we make one change to Equation 26: we remove the winner-denoising auxiliary by setting

$$\rho_{\mathrm{win}} = 0, \qquad a_{\mathrm{win},t} = 0, \qquad \mathcal{L}_{\mathrm{reg},t} = a_{\mathrm{ref},t}\mathcal{L}_{\mathrm{ref},t}.$$

This avoids pulling the SDXL denoiser toward preferred images as absolute targets when the pretrained model may already generate better images. The winner gate κ_t is therefore unnecessary; the reference term still limits drift at sampled noisy states. The resulting batch loss is

$$\widehat{\mathcal{L}}_{\mathrm{SDXL}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\Big[\mathrm{sg}(w_i)s_R\ell_h(R_{\theta,i}) + a_{\mathrm{ref},i}\mathcal{L}_{\mathrm{ref},i} + \lambda_{\mathrm{conf}}\ell_{\mathrm{conf},i}\Big]. \tag{86}$$

For the SDXL forward pass, we provide pooled text embeddings and image-specific size/crop conditioning to both the trainable and frozen-reference denoisers.

C.4 DMPO Baseline Implementation

We implement the SD1.5 DMPO baseline in Table 1 following the pairwise loss and pseudocode of Li et al. (2025). For each non-tied pair (c, x_0^w, x_0^l), we encode both images into latents, sample $t \in \{0, \ldots, 999\}$, and draw one Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. As in the paper's pseudocode, the two

branches share t and ϵ, giving $x_t^* = \alpha_t x_0^* + \sigma_t \epsilon$ for $* \in \{w, l\}$. On each branch, the trainable and frozen-reference denoisers receive identical (x_t^*, t, c). Define their four noise-prediction errors by

$$e_{M,t}^* = \|\epsilon - \epsilon_M(x_t^*, t, c)\|_2^2, \qquad M \in \{\theta, \text{ref}\}, \quad * \in \{w, l\}.$$

Each norm is summed over the latent coordinates of one image. The reference-relative error differences determine the logit u_t and the reverse-KL loss:

$$\begin{aligned} u_t &= -\beta_{\text{DMPO}}\left[(e_{\theta,t}^w - e_{\text{ref},t}^w) - (e_{\theta,t}^l - e_{\text{ref},t}^l)\right], \\ \ell_{\text{DMPO},t} &= \text{sigm}(u_t) \log \frac{\text{sigm}(u_t)}{1 - \alpha_{\text{DMPO}}} + \text{sigm}(-u_t) \log \frac{\text{sigm}(-u_t)}{\alpha_{\text{DMPO}}}. \end{aligned} \tag{87}$$

We average $\ell_{\text{DMPO},t}$ over pairs before backpropagation; reference predictions are detached, while gradients flow through the trainable denoiser. The target probability of the preferred image is $1 - \alpha_{\text{DMPO}}$, so the loss is $D_{\text{KL}}[\text{Bern}(\text{sigm}(u_t)) \,\|\, \text{Bern}(1 - \alpha_{\text{DMPO}})]$. We compute its log probabilities with stable log-sigmoid expressions. The paper absorbs the theoretical factor $T\omega(\lambda_t)$ into the effective β_{DMPO} used in its pseudocode.

For SD1.5, the paper reports $\alpha_{\text{DMPO}} = 0.01$ and $\beta_{\text{DMPO}} = 2000$. We retain the summed-error reduction above when using this β; a latent-coordinate mean would change the logit scale. Its training uses the 851,293 non-tied Pick-a-Pic v2 pairs from 58,960 prompts, four A100 40 GB GPUs, two pairs per GPU, 256 gradient-accumulation steps (effective batch size 2,048), a 10^{-5} learning rate, and 500 linear-warmup steps. It selects the best-performing checkpoint. The paper does not specify the optimizer, total training steps, or exact training resolution. We evaluate the DMPO baseline under the common SD1.5 generation and scoring protocol specified above.