TOPO: TOKEN-ORIENTED PREFERENCE OPTIMIZATION FOR TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **Qualitative comparison on four fixed SDXL prompts.** Columns show SDXL Base, Diffusion-DPO, InPO, MaPO, and ToPO. Orange bold prompt terms mark the target constraints.

ABSTRACT

Diffusion models have achieved remarkable success in text-to-image generation, yet aligning their outputs with human preferences remains challenging. Diffusion-DPO learns from image preference pairs, but their global labels do not specify how supervision should be allocated across spatial locations and denoising timesteps. The semantic roles of prompt tokens are also not explicitly considered when assigning these local weights. To address these limitations, we propose Token-Oriented Preference Optimization (ToPO), which builds a token-conditioned spatial–temporal weighting map for each pair. A frozen reference denoiser obtains spatial and temporal weights from local noise-prediction error differences under shared noise, while preferred-branch cross-attention links spatial evidence to content tokens. The map reweights Diffusion-DPO and supports an auxiliary pixel-midpoint ordering loss. It requires neither local preference annotations nor a separately trained reward model. Extensive experiments across multiple preference and compositional evaluations show that ToPO achieves state-of-the-art performance among the compared methods with only 500 training updates, demonstrating the potential of token-conditioned preference allocation for high-quality diffusion alignment within a limited update budget.

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

1 INTRODUCTION

Text-to-image diffusion models now generate high-fidelity images from natural-language prompts (Rombach et al., 2022). Yet their outputs still fall short of human preferences, both in visual quality and in faithfulness to the prompt, especially when a prompt specifies several objects, attributes, or spatial relations (Ghosh et al., 2023). Reinforcement learning and reward backpropagation can narrow this gap (Black et al., 2024; Prabhudesai et al., 2023), but they rely on a separately trained reward model and repeated online sampling. Diffusion-DPO (Wallace et al., 2024) removes both requirements by extending Direct Preference Optimization (Rafailov et al., 2023) to pairs of preferred and dispreferred images, and has become a standard basis for offline preference alignment (Li et al., 2024; Zhu et al., 2025; Kang et al., 2026).

However, this simplicity comes with a coarse view of supervision. Each pair contributes a single image-level comparison, and Diffusion-DPO spreads it evenly over every latent location and every denoising timestep. Yet the reason one image wins is rarely spread evenly, in either space or time. Consider two images generated for the prompt "a robot with black visor and number 42 on its chest" (Figure 1). One may win simply because the number on its chest is legible, a difference confined to a small region, while the two images also differ in background, pose, and lighting that the prompt never mentions. The decisive difference is confined in time as well: global layout and pose form at early, high-noise steps, whereas fine details such as the digits take shape at later steps (Balaji et al., 2022; Choi et al., 2022). Even allocation therefore spends much of each update on regions and timesteps that have little to do with why one image won, diluting the supervision carried by each pair. What each comparison lacks is an account of *where* and *when* it applies.

Recent work supplies this account, but draws it from outside the preference pair. Spatial methods assign credit to patches, instances, or regions (Huang et al., 2025; Sun et al., 2026; Liu et al., 2026b), relying on pretrained vision models, object-level annotations, or masks to decide which regions matter. Temporal methods score individual denoising steps with step-aware reward models trained on noisy latents (Liang et al., 2025; Zhang et al., 2025; Xian et al., 2026), which brings back the learned reward model that Diffusion-DPO was designed to remove. Each line also covers only one of the two axes, and allocating credit over both has so far relied on online reinforcement learning with a reward model (Shen et al., 2026). This raises a question: **can a preference pair itself tell *where* and *when* its supervision should go?**

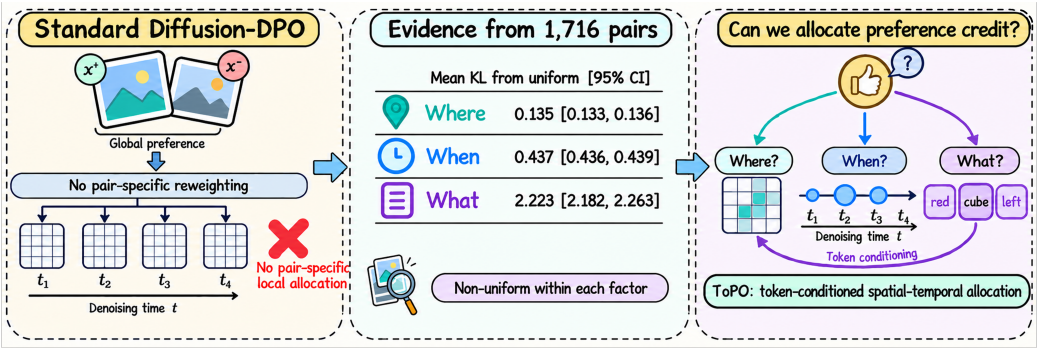

Figure 2: **Motivation for token-conditioned preference allocation.** Left: Diffusion-DPO spreads each image-level comparison evenly over latent locations and denoising timesteps. Middle: over 1,716 preference pairs, the weights derived from the reference error gap and cross-attention depart from uniform in *where*, *when*, and *what* (mean KL from uniform with 95% confidence intervals). Right: ToPO allocates each comparison by *where*, *when*, and *what* the pair disagrees about.

The answer can come from a component that Diffusion-DPO already has: its frozen reference denoiser. When the two images are corrupted with the same noise, any gap between the reference model's prediction errors on them comes from the images themselves. Measured at each location and timestep, this error gap shows where and when the model responds to the two images differently. The error gap does not tell which image is better, and it does not need to: the preference label already fixes the direction of the update, so the error gap only has to decide *where* and *when* the update lands. Across 1,716 Pick-a-Pic pairs, the weights it produces are clearly non-uniform in both space and time (Figure 1). The error gap, however, cannot tell which differences the prompt cares about. Two images sampled independently for the same prompt differ almost everywhere, so

the error gap highlights every difference the model responds to, whether or not the prompt mentions it. A change in background texture can weigh as much as the difference between a legible and an illegible "42".

Separating these differences requires knowing *what* each one is about, and the same reference denoiser already encodes this in its cross-attention, which ties each prompt token to the image regions that render it (Hertz et al., 2023; Chefer et al., 2023). Reading the error gap through these attention maps scores each content token by how much of the gap falls on the regions that attend to it. Because the scores come from the error gap, they favor the tokens on which the two images actually disagree, such as "42" in our example, rather than every token in the prompt. Projecting the scores back onto the image then raises the weight of regions tied to these tokens, while regions whose attention falls on no content token, such as the background texture, lose weight. In a controlled test, target color, shape, and texture tokens receive 26% to 40% more weight than matched non-target tokens (Appendix S5.1).

We turn these observations into Token-Oriented Preference Optimization (ToPO), which reweights the Diffusion-DPO loss by *where*, *when*, and *what* each pair disagrees about. ToPO has two components. **(1) Error-gap map.** The error gap of the frozen reference denoiser yields a weight for each latent location and a weight for each denoising timestep. **(2) Token conditioning.** Cross-attention on the preferred image, which renders the prompt in the way the rater favored, converts the spatial weights into token scores and projects them back, so that the spatial weights concentrate on differences tied to the prompt. In the loss, the spatial weights set how much each location contributes to the pairwise comparison, and the timestep weight sets how strongly the comparison is enforced at the sampled step. Because the weights come from a frozen model and carry no gradient, ToPO leaves the Diffusion-DPO pipeline unchanged apart from the weights inside its loss. It thus answers *where*, *when*, and *what* from the preference pair itself, with no local annotation and no reward model.

Our contributions are summarized as follows:

- We separate the direction of a preference update from its allocation. The pairwise label fixes the direction, and we show that the allocation over *where*, *when*, and *what* can be derived from the preference pair itself.
- We propose ToPO, a drop-in reweighting of Diffusion-DPO built from the error gap of its frozen reference denoiser and the cross-attention of the preferred image.
- Under a matched training budget on SD-1.5 and SDXL, ToPO improves over Diffusion-DPO on preference and compositional benchmarks. In a blind study on SDXL, raters prefer ToPO over Diffusion-DPO in 63.3% to 70.7% of comparisons. Controls that flatten or shuffle the weights indicate that both their non-uniformity and their placement matter.

2 Related Work

Preference alignment for diffusion models. DPO learns from paired preferences without an explicit reward model or an online reinforcement-learning loop (Rafailov et al., 2023). Diffusion alignment has explored reward-driven trajectory training (Black et al., 2024), pairwise objectives such as D3PO and Diffusion-DPO (Yang et al., 2024a; Wallace et al., 2024), per-image utility optimization in Diffusion-KTO (Li et al., 2024), and score-based preference learning (Zhu et al., 2025). Other work addresses noisy preference pairs and timestep or reference choices (Li et al., 2026b; Liu et al., 2026a; Kang et al., 2026). These methods mainly optimize image-level preferences and leave the allocation of supervision across local denoising coordinates unspecified.

Fine-grained spatial and instance alignment. Prior work has refined visual preference learning at different levels of granularity, from image regions to entire samples. PatchDPO introduces local patch supervision, while IAPO and BiDPO construct instance- or region-level preferences from image comparisons (Huang et al., 2025; Sun et al., 2026; Liu et al., 2026b). Visual preference optimization methods further provide pixel- or sample-level supervision, and OSPO uses object-centric data and attention masks (Ni et al., 2026; Li et al., 2026a; Oh et al., 2026). These methods demonstrate the value of localized credit, whereas ToPO derives a spatial route directly from the supplied preferred and dispreferred images.

Temporal and latent preference optimization. Trajectory-aware methods distribute preference supervision over denoising time or latent space through dense rewards, DDIM inversion, InPO, and related objectives (Yang et al., 2024b; Lu et al., 2025; Li et al., 2025; Wu et al., 2025). SPO, LPO,

and SLRM/TAPO further model step-aware or noise-compatible latent rewards (Liang et al., 2025; Zhang et al., 2025; Xian et al., 2026). These approaches motivate temporal credit assignment, while ToPO infers a separable time factor directly from frozen-reference residual contrast.

Attention-guided routing and control. Cross-attention connects text tokens with latent image features in diffusion models (Rombach et al., 2022). Prompt-to-Prompt and Attend-and-Excite use attention maps for inference-time editing and semantic control (Hertz et al., 2023; Chefer et al., 2023), while STAR studies adaptive spatiotemporal reward allocation for text-to-image reinforcement learning (Shen et al., 2026). ToPO uses preferred-branch cross-attention to convert pair-derived spatial evidence into token-conditioned DPO weights.

Taken together, prior work studies image-level preference optimization, spatial credit, temporal allocation, and attention control largely as separate problems. ToPO combines these factors in one offline objective by deriving a spatial-temporal route from frozen-reference residual evidence and conditioning it on content tokens through preferred-branch cross-attention.

3 Preliminaries: Diffusion Preference Learning

3.1 Conditional latent diffusion

Let x be an image, $z_0 = \mathcal{E}(x)$ its latent encoding, and c a text prompt. For noise $\varepsilon \sim \mathcal{N}(0, I)$, the DDPM forward process (Ho et al., 2020) produces

$$z_\tau = \sqrt{\bar{\alpha}_\tau} z_0 + \sqrt{1 - \bar{\alpha}_\tau}\, \varepsilon, \qquad \bar{\alpha}_\tau = \prod_{j \le \tau} (1 - \beta_j). \tag{1}$$

The denoiser $\varepsilon_\theta(z_\tau, \tau, c)$ predicts the added noise. Each prediction is indexed by a latent spatial location s and a denoising timestep τ. At each location and timestep, we define the local residual as

$$d_\theta(x; c, s, \tau) = \|\varepsilon_\theta(z_\tau, \tau, c)[s] - \varepsilon[s]\|_2^2 . \tag{2}$$

This residual is the basic quantity that a route can allocate across diffusion coordinates. Prompt tokens affect these coordinates through cross-attention; they do not index noise-prediction residuals directly.

3.2 Human preferences and Diffusion-DPO

The preference data consist of triples $Z = (c, x^+, x^-)$, where x^+ is preferred to x^- for prompt c. Under the Bradley–Terry model,

$$\Pr(x^+ \succ x^- \mid c) = \sigma\big(r^*(c, x^+) - r^*(c, x^-)\big)\,, \tag{3}$$

where r^* is an unobserved human reward. Direct preference optimization compares a learned policy score with a reference score, without fitting an explicit reward model (Rafailov et al., 2023). Diffusion-DPO instantiates this comparison with a tractable surrogate for the image-level diffusion likelihood (Wallace et al., 2024). Let $S_m(c, x)$ denote this surrogate for $m \in \{\theta, \mathrm{ref}\}$, and define $\Delta_m(Z) = S_m(c, x^+) - S_m(c, x^-)$. The pairwise objective is

$$\mathcal{L}_{\text{D-DPO}} = -\mathbb{E}_Z \log \sigma(\beta\, [\Delta_\theta(Z) - \Delta_{\mathrm{ref}}(Z)])\,. \tag{4}$$

The trajectory surrogate varies across diffusion preference objectives. In each case, a preference pair enters the loss through one image-level scalar comparison.

4 ToPO: Token-Oriented Preference Optimization

4.1 Overview

ToPO constructs a token-conditioned route over native diffusion coordinates from each preference pair. A frozen reference first evaluates both branches with shared noise at four anchors and channel-averages their local squared-residual contrast. Averaging this contrast over anchors yields the spatial factor $w_s(s)$; averaging it over latent sites yields discrete anchor-time evidence, which Gaussian interpolation turns into $w_\tau(\tau)$ at the sampled training timestep. Preferred-branch cross-attention reads the pre-prior spatial evidence into content-token weights $w_k(k)$ and projects their normalized, prior-mixed values back to the latent grid. This token-induced spatial map modulates $w_s(s)$ to form

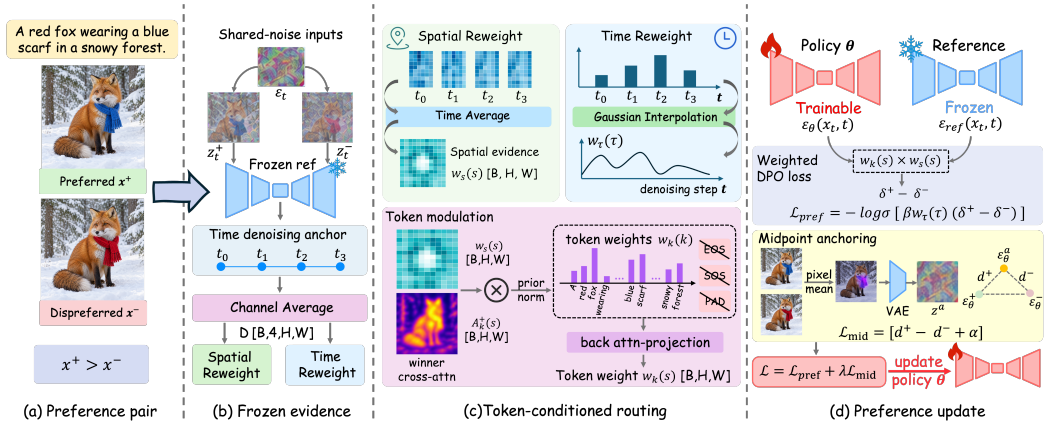

Figure 3: **ToPO routing pipeline.** A preferred–dispreferred pair is evaluated by a frozen reference under shared noise at four denoising anchors. Channel-averaged local residual contrasts are averaged over anchors to form spatial evidence and over latent sites to form discrete timestep evidence; normalized Gaussian interpolation evaluates the timestep factor at the sampled training step. Preferred-branch cross-attention reads the pre-prior spatial evidence into content-token weights. After content-token masking, prior mixing, and normalization, the token weights are backprojected through the same attention maps to modulate the spatial factor. The resulting detached spatial–temporal route reweights the Diffusion-DPO loss, and its spatial component also weights the pixel-midpoint ordering term. Pair labels specify the preferred direction; the route allocates the update over latent sites and denoising time.

the separable route $W(s,\tau)$. Thus, tokens modulate a diffusion-native route rather than defining a separate loss coordinate. The frozen reference constructs the detached route, whose spatial factor also weights the midpoint regularizer. Figure 3 summarizes the training-time computation.

4.2 Token-conditioned separable spatial–temporal route construction

For an anchor $\tau_i \in \{50, 349, 649, 949\}$, the residual-contrast pass noises both endpoints with the same draw ε_i and evaluates the frozen-reference predictions $\widehat{\varepsilon}^{\pm}_{\mathrm{ref},i} = \varepsilon_{\mathrm{ref}}(z^{\pm}_{\tau_i}, \tau_i, c)$. For channel q and latent site s, its branchwise squared-residual contrast is

$$R_i^{\pm}(q,s) = \omega_i\big(\varepsilon_i(q,s) - \widehat{\varepsilon}^{\pm}_{\mathrm{ref},i}(q,s)\big)^2, \qquad D_i(s) = \frac{1}{C}\sum_{q=1}^{C}\big|R_i^{+}(q,s) - R_i^{-}(q,s)\big|. \tag{5}$$

Here ω_i is the Diffusion-DPO integrand weight. $D_i(s)$ is a symmetric observable residual-contrast magnitude: it is high where the frozen reference exhibits different local residual magnitudes for the two image branches. We use this label-invariant quantity to assign *allocation mass*, while the signed winner–loser term $\delta^+ - \delta^-$ in equation 12 specifies preference direction. For example, a site with residual magnitudes $(0.1, 0.4)$ receives contrast 0.3 under either ordering of the branch labels. Thus the route emphasizes coordinates with larger branchwise residual-magnitude differences; it is not a standalone local preference label. Let Norm_1 normalize a nonnegative vector over its declared support to sum to one. The residual-contrast pass forms the data-dependent spatial and anchor-time marginals

$$d_s(s) = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} D_i(s), \qquad d_\tau(\tau_i) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} D_i(s), \tag{6}$$

then mixes each normalized marginal with its declared structural prior:

$$w_s = \mathrm{Norm}_1[\lambda\,\mathrm{Norm}_1[d_s] + (1-\lambda)p_s], \qquad \bar{w}_\tau(\tau_i) = \mathrm{Norm}_1[\lambda\,\mathrm{Norm}_1[d_\tau] + (1-\lambda)p_\tau]. \tag{7}$$

The locked configuration uses $\lambda = 0.5$, a uniform spatial prior, and an SNR-flat anchor-time prior. During training, the anchor-time distribution is evaluated at an arbitrary sampled timestep through

normalized Gaussian interpolation with $\sigma = 100$:

$$w_\tau(\tau) = |\mathcal{A}| \sum_{i \in \mathcal{A}} \kappa_\sigma(\tau - \tau_i)\, \bar{w}_\tau(\tau_i), \qquad \kappa_\sigma(\tau - \tau_i) = \frac{\exp(-|\tau - \tau_i|^2/(2\sigma^2))}{\sum_{j \in \mathcal{A}} \exp(-|\tau - \tau_j|^2/(2\sigma^2))}. \tag{8}$$

At every preferred-branch residual-contrast forward, we capture the attention probabilities of all U-Net modules named `attn2`. We average query–key attention probabilities over heads, average each captured layer over the four anchors, average the resized layer maps, and bilinearly resize the result to the native latent grid (64×64 for SD-1.5 and 128×128 for SDXL); denote the resulting map by $\bar{A}_k^+(s)$. We use the preferred branch as the declared coordinate-to-token readout: it conditions the spatial route on the image that defines the desired direction, while the signed DPO logit in equation 12 alone carries the winner–loser orientation. This choice is neither a target annotation nor a claim that preferred-only attention is the unique valid bridge; loser-only and symmetric bridges are distinct route variants. The token pathway reads the pre-prior spatial evidence $w_s^{\mathrm{data}} = \mathrm{Norm}_1[d_s]$ through this map:

$$w_k^{\mathrm{data}}(k) \;\propto \mathbf{1}_{\mathrm{content}}(k) \sum_s \bar{A}_k^+(s) w_s^{\mathrm{data}}(s), \qquad w_k = \mathrm{Norm}_1\left[\lambda\, \mathrm{Norm}_1[w_k^{\mathrm{data}}] + (1-\lambda) p_k\right]. \tag{9}$$

The content-token indicator removes padding, SOS, and the valid EOS position before normalization. This defines w_k as an attention-weighted token-conditioned modulation statistic of branchwise residual disagreement while retaining a content-token prior.

Cross-attention folds the factors into the field used by the objective:

$$\begin{aligned} W_s(s) &= \mathrm{Norm}_{\mathrm{mean}}\left(w_s(s) \left[\sum_k \bar{A}_k^+(s) w_k(k) \right] \right), \\ W(s,\tau) &= W_s(s) w_\tau(\tau), \qquad (\widetilde{W}_s, \widetilde{w}_\tau) = \mathrm{sg}[W_s, w_\tau], \\ \widetilde{W}(s,\tau) &= \widetilde{W}_s(s) \widetilde{w}_\tau(\tau). \end{aligned} \tag{10}$$

Here $w_s \in \mathbb{R}^{|\mathcal{S}|}$, $w_\tau \in \mathbb{R}^{|\mathcal{T}|}$, and $w_k \in \mathbb{R}^{|\mathcal{K}|}$, whereas $W \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{T}|}$. Equation 10 is deliberately separable over the two native loss coordinates: w_k modulates W_s, and does not introduce a three-dimensional (s, τ, k) loss tensor. At every update t, the frozen reference computes $W_t = \mathcal{W}_{\mathrm{ref}}(\mathcal{B}_t)$ from the current preference minibatch. The policy then uses $\widetilde{W}_t = \mathrm{sg}[W_t]$ on that same update: the route is differentiated through neither the attributor nor the reference model. This per-minibatch construction is shared by the SD-1.5 and SDXL implementations.

The normalization convention is part of the construction rather than a cosmetic rescaling. For a nonnegative map u on its declared support $\mathcal{J}$, we use

$$\mathrm{Norm}_{\mathrm{mean}}[u]_j = \frac{u_j}{\max\{|\mathcal{J}|^{-1} \sum_{j' \in \mathcal{J}} u_{j'},\, 10^{-8}\}}. \tag{11}$$

Consequently, multiplying the factors changes their arrangement across coordinates without allowing a vanishing or arbitrarily scaled factor to set the global loss scale. The 10^{-8} floor is the implementation safeguard used before the post-clipping update. This convention also makes the Shuffled-W stress test interpretable: it alters the arrangement of a normalized route rather than merely changing the total weight applied to the loss.

4.3 Preference and anchor losses

The main objective applies the detached route to the reference-relative local residual advantage. With policy and reference predictions $\widehat{\varepsilon}_\theta^{\pm}$ and $\widehat{\varepsilon}_{\mathrm{ref}}^{\pm}$ at a sampled training timestep, define

$$\begin{aligned} \delta^{\pm} &= \frac{1}{C|\mathcal{S}|} \sum_{q,s} \widetilde{W}_s(s) \left[\left(\widehat{\varepsilon}_{\mathrm{ref}}^{\pm}(q,s) - \varepsilon^{\pm}(q,s)\right)^2 - \left(\widehat{\varepsilon}_\theta^{\pm}(q,s) - \varepsilon^{\pm}(q,s)\right)^2 \right], \\ \mathcal{L}_{\mathrm{ToPO}} &= -\mathbb{E} \log \sigma\left(\beta \widetilde{w}_\tau(\tau)[\delta^+ - \delta^-]\right). \end{aligned} \tag{12}$$

Thus $\widetilde{W}_s$ weights the per-site residual reduction and $\widetilde{w}_\tau$ scales the sampled-timestep preference logit. We further form a deterministic pixel-midpoint anchor $x^a = \mathcal{E}((x^+ + x^-)/2)$ and penalize an

ordering violation between the policy's spatially weighted prediction distances to the two endpoints:

$$\mathcal{L}_{\text{tri}} = \mathbb{E}\left[\max\left\{0, \left\|(\widehat{\varepsilon}^{a}_{\theta} - \widehat{\varepsilon}^{+}_{\theta}) \odot \sqrt{\widetilde{W}_s}\right\|_2^2 - \left\|(\widehat{\varepsilon}^{a}_{\theta} - \widehat{\varepsilon}^{-}_{\theta}) \odot \sqrt{\widetilde{W}_s}\right\|_2^2 + \alpha\right\}\right]. \tag{13}$$

The complete loss is $\mathcal{L}_{\text{ToPO}} + \gamma\mathcal{L}_{\text{tri}}$, with $\gamma = 0.1$ and $\alpha = 0.05$ in the locked run configuration. We encode the pixel midpoint $\mathcal{E}((x^+ + x^-)/2)$ rather than interpolate the two endpoint latents; this fixes a deterministic, symmetric image-space anchor under the same encoder used for both endpoints. Pixel averaging can contain ghosted or otherwise off-manifold content, so the anchor is not treated as a natural-image target and no generated image is trained to imitate it. It only supplies a common reference for the relative ordering of the two spatially weighted prediction distances. Because its three predictions are evaluated at one sampled timestep, it uses the shared factor $\widetilde{W}_s$, not the full $W(s, \tau)$.

The controlled removals and scope of the supporting characterization are specified in Appendix S3. They keep the underlying preference pairs fixed and test sensitivity to the declared routing factors under the locked configuration.

5 Experiments

5.1 Datasets and implementation details

We train SD-1.5 and SDXL on Pick-a-Pic v2 (Kirstain et al., 2023). Removing approximately 12% tie comparisons following Diffusion-DPO (Wallace et al., 2024) leaves 851,293 ordered pairs from 58,960 prompts. SD-1.5 uses the released SFT-winners checkpoint at 512^2, while SDXL uses SDXL Base at 1024^2; in both cases, we finetune the policy U-Net and freeze the text encoder, VAE, and reference U-Net. All runs use eight NVIDIA H200 GPUs and 500 AdamW updates. We use a learning rate of 10^{-5}, a 200-step warmup, weight decay of 0.01, and gradient clipping at 1.0. Per-device batches of 4 and 2 with 32-step accumulation give effective batches of 1,024 and 512, respectively. SD-1.5 uses fp16 and SDXL bf16, with route construction and master weights in fp32. We use four route anchors, $\beta = 2000$, $\gamma = 0.1$, and margin 0.05; Appendix S3 gives the remaining settings.

5.2 Baselines and evaluation details

Baselines. For SD-1.5, we compare Base, Diffusion-DPO (Wallace et al., 2024), SFT-winners, KTO (Li et al., 2024), InPO (Lu et al., 2025), and ToPO; the SDXL comparison replaces KTO with MaPO (Hong et al., 2026). The multi-method tables use released checkpoints, while the three-seed ToPO–Diffusion-DPO study matches initialization, data order, optimizer, effective batch, and 500-update schedule (Appendix S4).

Benchmarks and metrics. On 500 Pick-a-Pic validation prompts, we report PickScore, HPSv2 (Wu et al., 2023), ImageReward (Xu et al., 2023), aesthetic score, and CLIP; PartiPrompts (1,632 prompts) and HPDv2 (3,200 prompts) results appear in Appendix S5. Compositional evaluation uses all 553 GenEval prompts (Ghosh et al., 2023) and 300 prompts per T2I-CompBench category (Huang et al., 2023).

User study. In the blind SDXL study, 16 consenting participants evaluate image pairs with the method names hidden. The study compares ToPO with Base and with Diffusion-DPO over 300 prompts, uses a shared seed, and allows ties. Participants judge overall preference, visual appeal, and prompt faithfulness. We report aggregate choice rates; Appendix S3.1 gives the full protocol.

5.3 Evaluation Results

Preference results. Table 1 compares final checkpoints on five automatic measures. On SD-1.5, ToPO ranks first on PickScore, HPSv2, ImageReward, and Aesthetic score. On SDXL, it ranks first on PickScore, HPSv2, ImageReward, and CLIP. Relative to Diffusion-DPO, ImageReward increases from 0.250 to 0.812 on SD-1.5 and from 0.934 to 1.032 on SDXL. HPSv2 also improves on both backbones. The matched three-seed comparison in Appendix S4 further confirms gains on all SD-1.5 measures and on SDXL HPSv2, ImageReward, and CLIP.

Compositional results. Table 2 evaluates whether the preference improvements extend to prompts with compositional constraints. The GenEval and T2I-CompBench averages are unweighted means over the displayed categories; the T2I-CompBench average is a manuscript-derived summary of

Table 1: Pick-a-Pic checkpoint results on 500 prompts for SD-1.5 and SDXL.

Backbone	Method	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑	CLIP ↑
SD-1.5	Base	20.543	0.2467	0.074	5.325	27.199
	Diffusion-DPO	20.947	0.2586	0.250	5.336	27.595
	SFT-winners	21.181	0.2824	0.642	5.354	28.064
	KTO	21.108	0.2810	0.618	5.351	28.003
	InPO	*21.410*	*0.2871*	*0.748*	*5.357*	**28.467**
	ToPO	**21.542**	**0.2903**	**0.812**	**5.361**	*28.393*
SDXL	Base	21.866	0.2869	0.744	5.359	28.411
	Diffusion-DPO	*22.264*	0.3005	0.934	*5.365*	*29.018*
	SFT-winners	21.481	0.2829	0.647	5.348	28.526
	InPO	22.208	*0.3044*	*0.955*	*5.365*	28.842
	MaPO	21.945	0.2962	0.866	**5.377**	28.613
	ToPO	**22.270**	**0.3140**	**1.032**	5.360	**29.487**

the six shared categories, not an official score. On SD-1.5, ToPO raises the GenEval average from 0.4591 to 0.5092 and the T2I-CompBench average from 0.3521 to 0.4123 relative to Diffusion-DPO. The gains cover the reported color, position, and shape categories. On SDXL, ToPO achieves the highest GenEval average and improves the reported GenEval components, while its T2I-CompBench average remains nearly tied with Diffusion-DPO.

Table 2: Compositional evaluation on GenEval and T2I-CompBench. Shading indicates point-estimate rank only.

	SD-1.5							SDXL						
	GenEval ↑			T2I-CompBench ↑				GenEval ↑			T2I-CompBench ↑			
Method	Avg.‡	Colors	Position	Avg.†	Color	Shape	Method	Avg.‡	Colors	Position	Avg.†	Color	Shape
Base	0.4378	0.7553	0.0500	0.3393	0.3750	0.3715	Base	0.5783	0.8723	*0.1275*	0.4533	0.6277	0.4878
Diffusion-DPO	0.4591	0.7926	0.0625	0.3521	0.4028	0.3842	Diffusion-DPO	*0.5917*	0.8644	0.1075	*0.4974*	*0.7177*	*0.5378*
SFT-winners	0.4878	0.8138	*0.0725*	0.3870	0.4761	0.4310	SFT-winners	0.5464	0.8670	0.1250	0.4436	0.6089	0.4955
KTO	0.4879	*0.8271*	0.0550	0.3887	0.4875	0.4300	InPO	0.5827	0.8750	0.1125	0.4731	0.6731	0.5231
InPO	*0.4925*	0.8245	*0.0725*	*0.3984*	*0.4967*	*0.4389*	MaPO	0.5687	*0.8803*	0.1225	0.4691	0.6429	0.5197
ToPO	**0.5092**	**0.8404**	**0.0775**	**0.4123**	**0.5242**	**0.4579**	ToPO	**0.6168**	**0.9069**	**0.1675**	**0.4975**	**0.7190**	**0.5398**

Human evaluation. In the blind SDXL study, 16 participants compare image pairs with the method names hidden, and ties are allowed. Figure 4 shows that ToPO receives the majority of votes for overall preference, visual appeal, and prompt faithfulness. Its win rates are 66.0%, 73.3%, and 67.3% against Base, and 63.3%, 70.7%, and 65.3% against Diffusion-DPO, respectively.

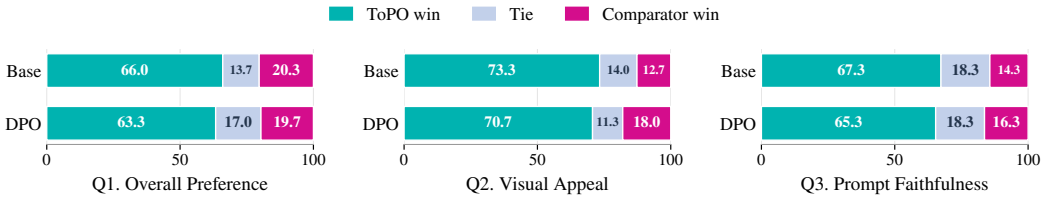

Figure 4: **Blind SDXL preference study.** Aggregate choice rates for overall preference, visual appeal, and prompt faithfulness in comparisons of ToPO with Base and Diffusion-DPO. Bars separate ToPO wins, ties, and comparator wins.

5.4 Ablation Studies and Controlled Route Analysis

We use Figure 5 and Table 3 to examine training length and the contributions of the route design under the locked SD-1.5 configuration.

Training-step ablation. Figure 5 traces one matched SD-1.5 run; the gray dashed lines show the author-released Diffusion-DPO checkpoint scores reported in Table 1. Within the matched run, ToPO surpasses Diffusion-DPO early and reaches near-best observed values on most measures by around 300 updates. It remains higher on PickScore, HPSv2, and ImageReward through 500 updates, suggesting that strong alignment can be reached before the end of the fixed schedule.

Route and component ablations. Table 3 jointly examines route allocation, individual route components, and preference-scale sensitivity. Uniform-W replaces the learned route with a uniform

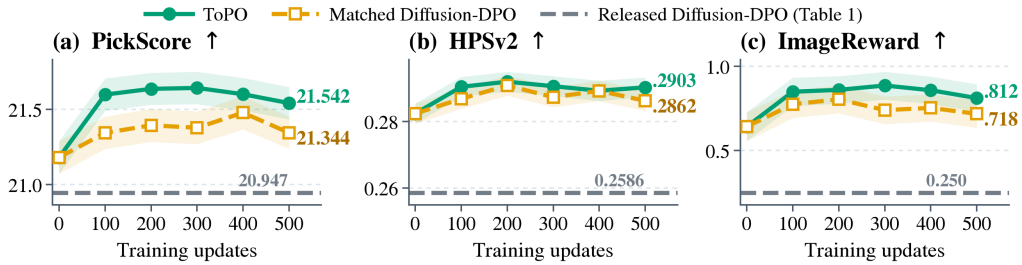

Figure 5: **Controlled SD-1.5 training dynamics.** PickScore, HPSv2, and ImageReward are evaluated at each checkpoint. Bands are 95% prompt-bootstrap intervals on 500 prompts; seedwise endpoints are in Appendix S4. Gray dashed lines show the author-released Diffusion-DPO checkpoint scores from Table 1, as fixed references distinct from the matched training trajectory.

weight while retaining the midpoint term, removing structured allocation across spatial locations, denoising steps, and tokens. Shuffled-W preserves the materialized route values but randomly changes their assignment to the corresponding coordinates. The full route obtains the largest point estimate across the displayed outcomes, indicating that the gains depend on where the weights are applied rather than only on their overall magnitude. Setting the spatial, timestep, or token factor to one lowers PickScore by 0.295, 0.307, and 0.168, respectively, showing that all three factors contribute to the final alignment behavior. Removing SOS/EOS masking causes a larger decrease of 0.429, confirming that special-token attention can distort token normalization. The midpoint term improves both PickScore and GenEval, while the preference-scale sweep is non-monotonic across metrics and identifies $\beta = 2000$ as the strongest overall setting.

Table 3: SD-1.5 ToPO ablations and route-allocation controls. Panel A uses one locked run per route allocation. Shading marks point-estimate rank, not significance.

Configuration	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑	CLIP ↑	GE Avg.‡ ↑	GE Count ↑	Final r-gap
A. Route-allocation controls (one locked run per configuration)								
ToPO (full)	**21.542**	**.2903**	**.812**	**5.361**	**28.393**	**.5092**	**.4969**	3.64×10^{-4}
Uniform-W	21.237	.2851	*.769*	5.348	*28.198*	.4710	*.4418*	4.29×10^{-4}
Shuffled-W	*21.382*	*.2899*	.741	*5.358*	28.180	*.4926*	.4375	-3.47×10^{-5}
B. Component removals								
ToPO (full)	**21.542**	**.2903**	**.812**	5.361	*28.393*	**.5092**	**.4969**	3.64×10^{-4}
$w_k \equiv 1$	*21.374*	.2842	.803	**5.365**	28.271	.5026	.4406	5.53×10^{-4}
$w_\tau \equiv 1$	21.235	*.2883*	.805	5.361	28.331	*.5079*	.4500	4.27×10^{-4}
$w_s \equiv 1$	21.247	.2856	*.808*	*5.364*	28.371	.5026	.4219	5.34×10^{-4}
No SOS/EOS prior	21.113	.2845	.788	5.362	28.311	.5060	*.4531*	6.50×10^{-4}
No triplet term	21.295	.2840	.807	5.363	**28.412**	.4950	.4375	6.81×10^{-4}
C. Preference scale β								
500	*21.484*	.2902	.804	**5.368**	28.277	.4894	.4062	1.00×10^{-3}
1000	21.433	**.2921**	**.856**	*5.364*	28.201	.4901	*.4406*	8.64×10^{-4}
2000 (default)	**21.542**	*.2903*	*.812*	5.361	**28.393**	**.5092**	**.4969**	3.64×10^{-4}
4000	21.398	.2883	.774	5.362	*28.298*	*.4972*	.4312	4.85×10^{-4}

6 Conclusion

We introduce ToPO, a token-conditioned preference optimization method for diffusion models that uses preferred-branch cross-attention to relate prompt tokens to spatial residual evidence and translates each global preference comparison into reweighted local noise-prediction updates across spatial locations and denoising timesteps. At four denoising anchors, a frozen reference denoiser measures branchwise residual contrast to construct spatial and timestep evidence without local preference labels or a learned reward model. Cross-attention reads the pre-prior spatial evidence into content-token weights and projects the normalized token weights back to the latent grid, refining the spatial route. The resulting detached spatial–temporal route reweights the Diffusion-DPO objective, while pixel-midpoint anchoring supplies a symmetric auxiliary ordering constraint. Extensive experiments across preference and compositional evaluations demonstrate the effectiveness of ToPO. ToPO provides a new perspective on diffusion preference learning by routing global preference feedback through diffusion's native spatial and temporal coordinates while using prompt tokens to identify the image content that this feedback should emphasize.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. URL https://arxiv.org/abs/2211.01324.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, pp. 4965–4987, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/14f75513f0f1ca01de1e826b52e6b840-Abstract-Conference.html.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, 2023. doi: 10.1145/3592116. URL https://arxiv.org/abs/2301.13826.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Choi_Perception_Prioritized_Training_of_Diffusion_Models_CVPR_2022_paper.html.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. doi: 10.52202/075280-2270. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

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://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4744–4752, 2026. doi: 10.1609/aaai.v40i6.42476. URL https://ojs.aaai.org/index.php/AAAI/article/view/42476.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023. doi: 10.52202/075280-3443.

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, and Jie Song. PatchDPO: Patch-level DPO for finetuning-free personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18369–18378, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_PatchDPO_Patch-level_DPO_for_Finetuning-free_Personalized_Image_Generation_CVPR_2025_paper.html.

Junyong Kang, Seohyun Lim, Kyungjune Baek, and Hyunjung Shim. Rethinking direct preference optimization in diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5611–5619, 2026. doi: 10.1609/aaai.v40i7.37480. URL https://ojs.aaai.org/index.php/AAAI/article/view/37480.

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, pp. 36652–36663, 2023. doi: 10.52202/075280-1594.

Ming Li, Jie Wu, Jiaxing Cui, Xiaojie Li, Rui Wang, and Chen Chen. ViPO: Visual preference optimization at scale. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/33bb58be3f0e903c75afa73d75b5c67e-Abstract-Conference.html.

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, pp. 24897–24925, 2024. doi: 10.52202/079017-0785. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2c487f8a54cf24c0684c32abc77fed56-Abstract-Conference.html.

Yang Li, Songlin Yang, Wei Wang, Xiaoxuan Han, and Jing Dong. α-dpo: Robust preference alignment for diffusion models via α divergence. In *International Conference on Learning Representations*, pp. 101046–101074, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a4c17d9b88eaefc9bdf7c656ffc8ce55-Abstract-Conference.html.

Zejian Li, Yize Li, Chenye Meng, Zhongni Liu, Ling Yang, Shengyuan Zhang, Guang Yang, Changyuan Yang, Zhiyuan Yang, and Lingyun Sun. Inversion-DPO: Precise and efficient post-training for diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9901–9910, 2025. doi: 10.1145/3746027.3755220. URL https://doi.org/10.1145/3746027.3755220.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13208, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_Aesthetic_Post-Training_Diffusion_Models_from_Generic_Preferences_with_Step-by-step_Preference_CVPR_2025_paper.html.

Xinxin Liu, Ming Li, Zonglin Lyu, Yuzhang Shang, and Chen Chen. Learning from noisy preferences: A semi-supervised learning approach to direct preference optimization. In *International Conference on Learning Representations*, pp. 13690–13709, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17061a94c3c7fda5fa24bbdd1832fa99-Abstract-Conference.html.

Zhuohan Liu, Wujian Peng, Yitong Chen, and Zuxuan Wu. Compositional text-to-image generation via region-aware bimodal direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36604–36614, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Compositional_Text-to-Image_Generation_Via_Region-aware_Bimodal_Direct_Preference_Optimization_CVPR_2026_paper.html.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. InPO: Inversion preference optimization with reparametrized DDIM for efficient diffusion model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28629–28639, 2025.

Ziqi Ni, Yuanzhi Liang, Rui Li, Yi Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. Seeing what matters: Visual preference policy optimization for visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27260–27269, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Ni_Seeing_What_Matters_Visual_Preference_Policy_Optimization_for_Visual_Generation_CVPR_2026_paper.pdf.

Yoonjin Oh, Yongjin Kim, Hyomin Kim, Donghwan Chi, and Sungwoong Kim. OSPO: Object-centric self-improving preference optimization for text-to-image generation. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7620–7631, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Oh_OSPO_Object-Centric_Self-Improving_Preference_Optimization_for_Text-to-Image_Generation_CVPR_2026_paper.html.

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/2310.03739v2. Withdrawn; superseded by arXiv:2407.08737.

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, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

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.

Jinjie Shen, Wei Deng, Xian Hu, Daiguo Zhou, and Jian Luan. STAR: Spatiotemporal adaptive reward allocation for text-to-image RL post-training. *arXiv preprint arXiv:2606.17979*, 2026. URL https://arxiv.org/abs/2606.17979.

Jiayang Sun, Pin Wang, Hongbo Wang, Xinyue Liu, Huaibo Huang, and Ran He. Towards fine-grained attribution: Instance-aware preference optimization for aligning diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43155–43164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_Towards_Fine-Grained_Attribution_Instance-Aware_Preference_Optimization_for_Aligning_Diffusion_Models_CVPR_2026_paper.html.

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.

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. *arXiv preprint arXiv:2306.09341*, 2023.

Xun Wu, Shaohan Huang, Lingjie Jiang, and Furu Wei. Rethinking DPO-style diffusion aligning frameworks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18068–18077, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_Rethinking_DPO-style_Diffusion_Aligning_Frameworks_ICCV_2025_paper.html.

Xiaole Xian, Xilin He, Wenting Chen, Wenshuang Liu, Wenqi Mu, Yancheng He, Liang Li, Yi Zhang, and Xiangyu Yue. Consistent noisy latent rewards for trajectory preference optimization in diffusion models. In *International Conference on Learning Representations*, pp. 6315–6340, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.html.

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, pp. 15903–15935, 2023. doi: 10.52202/075280-0700.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Using_Human_Feedback_to_Fine-tune_Diffusion_Models_without_Any_Reward_CVPR_2024_paper.html.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55998–56032, 2024b. URL https://proceedings.mlr.press/v235/yang24e.html.

Tao Zhang, Cheng Da, Kun Ding, Huan Yang, Kun Jin, Yan Li, Tingting Gao, Di Zhang, Shiming Xiang, and Chunhong Pan. Diffusion model as a noise-aware latent reward model for step-level preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106202–106235, 2025. doi: 10.52202/085713-3545. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/98d112a7111970b1638dafc97d96d9ee-Abstract-Conference.html.

Huaisheng Zhu, Teng Xiao, and Vasant Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *International Conference on Learning Representations*, pp. 51112–51133, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7f70331dbe58ad59d83941dfa7d975aa-Abstract-Conference.html.

Appendix

S1 AI USE STATEMENT

No generative AI tools were used to formulate the research questions or hypotheses, develop the methodology, design or conduct the experiments, generate or process data, analyze the results, or prepare the figures and tables reported in this paper. The scientific claims and conclusions were developed and verified by the authors, who take full responsibility for the content of the submission. A generative AI assistant was used solely to help draft the wording of this disclosure statement; the authors reviewed and finalized that wording.

S2 SUPPORTING ANALYSIS

This section provides supporting analyses for the detached route used by ToPO. The first analysis models the route as a current-minibatch field whose weights are recomputed by the frozen reference and detached before policy differentiation. The second establishes consistency of the attention-weighted token statistic when evaluated at a fixed checkpoint on held-out samples. We also give a simple scale bound for the routed preference logit.

S2.1 DETACHED ROUTE FIELD

Let Z_{t+1} contain a preference minibatch together with the diffusion randomness used to evaluate its loss, and let $\mathcal{F}_t$ denote the optimization history through θ_t. Because the route is recomputed by the frozen reference and detached before policy differentiation, define

$$\widetilde{G}_t = \nabla_\theta^{\mathrm{sg}} \ell(\theta_t, Z_{t+1}; \mathcal{W}_{\mathrm{ref}}(Z_{t+1})), \qquad G_t = \mathrm{Clip}_{1.0}(\widetilde{G}_t), \qquad h(\theta) = \mathbb{E}_{Z\sim P_0}[G(\theta, Z)]. \tag{14}$$

We consider a projected-SGD surrogate with independently resampled minibatches, bounded post-clipping gradients, local dissipativity around a stable root $\theta^\dagger$, and a Robbins–Monro step-size schedule for the almost-sure convergence statement. These assumptions describe the detached route field. The experiments themselves use constant-step AdamW.

Proposition 1 (Current-minibatch detached field)**.** *Under the stated surrogate assumptions, the noise term*

$$\xi_{t+1} = G_t - h(\theta_t)$$

is a martingale difference, and the corresponding projected iteration converges almost surely to $\theta^\dagger$.

Indeed, conditional on $\mathcal{F}_t$, the same current draw is used to recompute $\mathcal{W}_{\mathrm{ref}}(Z_{t+1})$. Therefore,

$$\mathbb{E}[G(\theta_t, Z_{t+1}) \mid \mathcal{F}_t] = h(\theta_t). \tag{15}$$

Let $e_t = \theta_t - \theta^\dagger$. Projection nonexpansiveness and the post-clipping gradient bound yield

$$\mathbb{E}\Big[\|e_{t+1}\|_2^2 \mid \mathcal{F}_t\Big] \le (1 - 2\mu\eta_{t+1}) \|e_t\|_2^2 + \eta_{t+1}^2 G_0^2. \tag{16}$$

A standard summable-drift argument then gives the stated convergence result. For a constant step size $0 < \eta < 1/(2\mu)$, the same recursion gives the following neighborhood bound.

Proposition 2 (Gradient-energy and constant-step bounds)**.** *Suppose that the unclipped local gradient can be written as*

$$\widetilde{G}_t = RD(W_t)\mathbf{u}_t.$$

Then

$$\mathrm{Var}(G_t \mid \mathcal{F}_t) \le \|R\|_{\mathrm{op}}^2 \, \mathbb{E}\Big[\|W_t\|_\infty^2 \|\mathbf{u}_t\|_2^2 \mid \mathcal{F}_t\Big] - \|h(\theta_t)\|_2^2. \tag{17}$$

Moreover,

$$\mathbb{E}\Big[\big\|\theta_T - \theta^\dagger\big\|_2^2\Big] \le (1 - 2\mu\eta)^T \big\|\theta_0 - \theta^\dagger\big\|_2^2 + \frac{\eta G_0^2}{2\mu}. \tag{18}$$

The first inequality follows from

$$\|RD(W_t)\mathbf{u}_t\|_2 \le \|R\|_{\mathrm{op}} \|W_t\|_\infty \|\mathbf{u}_t\|_2$$

together with clipping. The second follows by unrolling equation 16. It characterizes the neighborhood induced by a constant-step SGD surrogate and does not model the moment dynamics of AdamW.

S2.2 Attention-weighted residual contrast

Fix a checkpoint $\bar{\theta}$ and evaluate the detached attributor on held-out independent samples. Let

$$M_{\bar{\theta}}(Z, s) = \mathrm{Norm}_1[d_s(Z, s)]$$

denote the pre-prior spatial residual contrast, and let $\bar{A}^+_{k,\bar{\theta}}(s; Z)$ denote the aggregated winner-side attention mass for token k. The corresponding population token statistic is

$$\rho_k^{\mathrm{disc}} = \frac{\mathbb{E}_{Z\sim P_0}\left[\sum_{s\in\mathcal{S}} \bar{A}^+_{k,\bar{\theta}}(s; Z) M_{\bar{\theta}}(Z, s)\right]}{\mathbb{E}_{Z\sim P_0}\left[\sum_{s\in\mathcal{S}} \bar{A}^+_{k,\bar{\theta}}(s; Z)\right]}. \tag{19}$$

Proposition 3 (Consistency of the frozen token statistic)**.** *For a fixed finite spatial grid, independent held-out samples, integrable numerator and denominator terms, and positive expected attention mass, the self-normalized pre-prior statistic $\widehat{w}^{\mathrm{data}}_{k,n}$ converges in probability to ρ_k^{disc}.*

For sample Z_i, define

$$X_{i,k} = \sum_s \bar{A}^+_{k,\bar{\theta}}(s; Z_i) M_{\bar{\theta}}(Z_i, s), \qquad Y_{i,k} = \sum_s \bar{A}^+_{k,\bar{\theta}}(s; Z_i).$$

The empirical statistic is

$$\widehat{w}^{\mathrm{data}}_{k,n} = \frac{\sum_i X_{i,k}}{\sum_i Y_{i,k}}.$$

The weak law of large numbers gives convergence of the numerator and denominator averages. Since $\mathbb{E}[Y_{i,k}] > 0$, the continuous-mapping theorem yields the stated convergence.

S2.3 Routed-logit scale

The effect of the preference scale can be bounded directly for both a local routed contribution and an aggregate routed logit.

Proposition 4 (β–W scale envelope)**.** *For a local routed contribution $\beta W_j \delta_j$ and an aggregate routed logit, respectively,*

$$|\beta W_j \delta_j| \le \beta \|W\|_\infty |\delta_j|, \qquad |\beta \langle W, \boldsymbol{\delta}\rangle| \le \beta \|W\|_\infty \|\boldsymbol{\delta}\|_1 . \tag{20}$$

These bounds follow from $|W_j| \le \|W\|_\infty$ and the triangle inequality. They motivate monitoring $\beta \|W\|_\infty$ together with the finite preference-scale sweep reported in Table 3.

S3 Training and Evaluation Protocol

The SD-1.5 endpoint is obtained by full U-Net finetuning from the SFT-winners checkpoint at 512^2, while the SDXL endpoint is obtained by full U-Net finetuning from SDXL Base at 1024^2. Both models are trained for 500 AdamW updates with learning rate 10^{-5}, a 200-update linear warmup, weight decay 0.01, gradient clipping at 1.0, gradient checkpointing, and checkpoint evaluation every 100 updates. SD-1.5 uses a per-device batch size of 4 and SDXL uses a per-device batch size of 2. With 32 accumulation steps over eight processes, the effective batch sizes are 1,024 and 512, respectively.

Master U-Net weights and route construction use fp32, while policy autocasting uses fp16 for SD-1.5 and bf16 for SDXL. The frozen reference aggregates residual contrast and preferred-branch

cross-attention over the declared heads, layers, and four denoising anchors. Padding, SOS, and EOS tokens are excluded before token normalization. Algorithm 1 specifies the order of one minibatch update: the route is reconstructed from the current preference pairs and detached before the policy backward pass.

Algorithm 1 One detached preference-update step for ToPO.

Require: Frozen reference attributor; policy θ; current preference minibatch $\mathcal{B}_t$
Fixed settings: $\mathcal{A} = \{50, 349, 649, 949\}$; $\beta = 2000$, $\gamma = 0.1$, $\alpha = 0.05$
Run the residual-contrast route on $\mathcal{B}_t$ in full precision at each $\tau_i \in \mathcal{A}$; collect $\{D_i, A_i^+\}_{i\in\mathcal{A}}$
Construct, mask, and normalize w_s, w_τ, and w_k using Eqs. (5)–(9)
$W_s(s) \leftarrow \mathrm{Norm}_{\mathrm{mean}}\big(w_s(s)\sum_k \bar{A}_k^+(s) w_k(k)\big)$
$W(s,\tau) \leftarrow W_s(s) w_\tau(\tau)$
$(\widetilde{W}_s, \widetilde{w}_\tau) \leftarrow \mathrm{sg}[W_s, w_\tau]$ ▷ detach route
$\widetilde{W}(s,\tau) \leftarrow \widetilde{W}_s(s)\widetilde{w}_\tau(\tau)$
Draw diffusion noise and training times for $\mathcal{B}_t$; evaluate $\mathcal{L}_{\mathrm{ToPO}}$ with $\widetilde{W}$
$x^a \leftarrow \mathcal{E}((x^+ + x^-)/2)$ for each pair; evaluate $\mathcal{L}_{\mathrm{tri}}$ with $\widetilde{W}_s$ ▷ Eq. (13)
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{ToPO}} + \gamma\mathcal{L}_{\mathrm{tri}}$; compute $G \leftarrow \nabla_\theta \mathcal{L}$
$G \leftarrow \mathrm{Clip}_{1.0}(G)$; update $\theta \leftarrow \mathrm{AdamW}(\theta, G)$ with the locked learning-rate schedule
return updated θ

S3.1 Evaluation protocol

Table 4 specifies the generation and scoring settings used for each benchmark. Within each backbone and benchmark block, all methods use the same sampling and evaluation schedule, while SD-1.5 and SDXL results are reported separately. GenEval and T2I-CompBench use their official prompt files without subsampling.

Table 4: **Locked automatic-evaluation protocol.** Generation is benchmark-specific and held fixed within each backbone–benchmark comparison. Main and matched-seed endpoints use the final 500-update checkpoint; the external checkpoint rows retain their released endpoint. "No extra negative prompt" means that no additional negative-prompt argument is supplied to the generation call.

Protocol	Prompt / image schedule	Generation	Scorers / frozen implementation
Pick-a-Pic validation	500 prompts; one image per prompt; seed 0	DDIMScheduler from native pipeline config; 30 steps, CFG 7.5, no extra negative prompt; native VAE; 512^2 (SD-1.5) / 1024^2 (SDXL)	PickScore v1; HPS-v2 1.2.0 (v2.1 weights); ImageReward-v1.0; improved aesthetic predictor; CLIP ViT-L/14
T2I-CompBench	300 prompts per declared category; SD-1.5: 10 images/prompt with seeds 42–51; SDXL: 5 images/prompt with seeds 42–46	DDIMScheduler from native pipeline config; 30 steps, CFG 7.5, no extra negative prompt; native VAE; 512^2 / 1024^2	Official T2I-CompBench evaluator, commit `1b709499`; reported SDXL categories exclude the 10-image complex evaluator
GenEval	553 official prompts; four sequential samples/prompt from seed 42	DDIMScheduler from native pipeline config; 50 steps, CFG 9.0, no extra negative prompt; native VAE; 512^2 / 1024^2	Official GenEval scoring with the mmdetection-v3.3.0 adapter

S4 Matched Control Studies

For each backbone and seed, ToPO and Diffusion-DPO share the initial checkpoint, 500-update budget, preference-stream order, effective batch size, optimizer, learning-rate schedule, and $\beta = 2000$. All results are evaluated at the final 500-update checkpoint.

Table 5: **Matched three-seed comparison with Diffusion-DPO.** Mean ± sample standard deviation over three locked $\beta = 2000$ training-stream pairs, evaluated at the final 500-update checkpoint on the fixed 500-prompt protocol. Δ is computed for each seed as ToPO minus Diffusion-DPO. The SD-1.5 and SDXL blocks use separate protocols.

Backbone	Method / effect	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑	CLIP ↑
SD-1.5	Diffusion-DPO	20.930 ± 0.057	0.2587 ± 0.0006	0.250 ± 0.001	5.340 ± 0.012	27.609 ± 0.070
	ToPO	21.549 ± 0.059	0.2895 ± 0.0014	0.813 ± 0.015	5.359 ± 0.013	28.365 ± 0.032
	Paired Δ (ToPO–DPO)	**+0.620 ± 0.096**	**+0.0307 ± 0.0019**	**+0.563 ± 0.016**	**+0.019 ± 0.006**	**+0.756 ± 0.038**
SDXL	Diffusion-DPO	22.285 ± 0.053	0.3003 ± 0.0007	0.932 ± 0.000	5.370 ± 0.012	28.992 ± 0.066
	ToPO	22.253 ± 0.030	0.3135 ± 0.0020	1.030 ± 0.020	5.355 ± 0.007	29.499 ± 0.109
	Paired Δ (ToPO–DPO)	**-0.032 ± 0.026**	**+0.0132 ± 0.0016**	**+0.098 ± 0.020**	**-0.015 ± 0.012**	**+0.507 ± 0.159**

On SD-1.5, all five paired means favor ToPO. On SDXL, the paired means favor ToPO on HPSv2, ImageReward, and CLIP, while Diffusion-DPO is higher on PickScore and Aesthetic. The table reports the mean and sample standard deviation across the three matched training pairs.

S5 Extended Results and Qualitative Evidence

The main-text comparison evaluates Pick-a-Pic on 500 prompts. Here, we extend the evaluation of ToPO and the within-backbone baselines to GenEval, T2I-CompBench, PartiPrompts, and HPDv2 under the locked protocol. GenEval includes 553 prompts, T2I-CompBench uses 300 prompts per category, and the held-out PartiPrompts and HPDv2 suites contain 1,632 and 3,200 prompts, respectively. The qualitative figures below complement these numerical comparisons.

Table 6 reports the GenEval components, and Table 7 reports held-out PartiPrompts scores. Table 8 expands the T2I-CompBench results by category, while Table 9 reports HPDv2 scores. Table 10 presents a separate frozen-reference token-modulation diagnostic on 300 controlled cases.

S5.1 Token-Modulation and Scale Diagnostics

The fixed-template controlled token test contains 300 cases, with 60 cases each for color, shape, texture, count, and spatial relations. It compares the target-token w_k mass with matched non-target content-token controls using class-wise paired analyses. Table 10 reports this diagnostic for the frozen `sft_init_ref_unet` reference attributor. It therefore characterizes the attribution rule used by the diagnostic rather than the localization behavior of the final trained checkpoint.

A single-pair view of token-conditioned spatial–temporal routing. The controlled token test characterizes a fixed reference attributor over a prompt suite. Figure 11 instead shows a final-checkpoint diagnostic for one held-out SD-1.5 Pick-a-Pic v2 preference pair. In this example, the spatial factor is nonuniform, the timestep factor assigns different weights to the denoising anchors, and the attention-weighted token readout projects content-token modulation back to the latent grid. The resulting $W(s, \tau)$ is therefore a nonuniform spatial–temporal multiplier for this pair.

S6 Scope and Limitations

ToPO is designed for attention-equipped latent diffusion models that predict noise and expose both local residuals and cross-attention maps. Applying the method to flow-matching models or other conditioning interfaces would require a compatible allocation statistic and dedicated evaluation.

The matched three-seed study covers only the locked SD-1.5 and SDXL configurations, while comparisons with released checkpoints are descriptive rather than matched retraining experiments. The Shuffled-W control examines the effect of coordinate assignment but does not establish superiority under a matched compute budget, semantic localization, or an exhaustive comparison with alternative saliency constructions. The token diagnostic is computed with a frozen SD-1.5 reference

Table 6: Expanded GenEval results. GE Avg.‡ is the unweighted mean of the six displayed GenEval components.

Backbone	Method	GE Avg.‡ ↑	Single ↑	Two ↑	Count ↑	Colors ↑	Position ↑	Color-attr. ↑
SD-1.5	Base	0.4378	0.9781	0.4116	0.3719	0.7553	0.0500	0.0600
	Diffusion-DPO	0.4591	0.9812	0.4343	0.3938	0.7926	0.0625	0.0900
	SFT-winners	0.4878	**0.9875**	*0.5404*	0.4125	0.8138	*0.0725*	0.1000
	KTO	0.4879	**0.9875**	0.5202	*0.4375*	*0.8271*	0.0550	0.1000
	InPO	*0.4925*	*0.9844*	0.5379	0.4281	0.8245	*0.0725*	**0.1075**
	ToPO	**0.5092**	**0.9875**	**0.5480**	**0.4969**	**0.8404**	**0.0775**	*0.1050*
SDXL	Base	0.5783	*0.9906*	0.7955	0.4188	0.8723	*0.1275*	0.2650
	Diffusion-DPO	*0.5917*	0.9875	**0.8434**	0.4500	0.8644	0.1075	**0.2975**
	SFT-winners	0.5464	*0.9906*	0.7273	0.3562	0.8670	0.1250	0.2125
	InPO	0.5827	0.9781	0.8106	*0.4750*	0.8750	0.1125	0.2450
	MaPO	0.5687	0.9875	0.7904	0.3812	*0.8803*	0.1225	0.2500
	ToPO	**0.6168**	**1.0000**	*0.8409*	**0.4920**	**0.9069**	**0.1675**	*0.2935*

Table 7: Held-out PartiPrompts results on 1,632 prompts using one random training seed.

Backbone	Method	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑	CLIP ↑
SD-1.5	Base	21.275	0.2513	0.205	5.318	27.103
	Diffusion-DPO	21.509	0.2592	0.359	5.328	27.373
	SFT-winners	21.626	0.2804	0.653	5.354	27.754
	KTO	21.611	0.2795	0.618	5.351	27.734
	InPO	*21.785*	*0.2839*	*0.736*	*5.355*	*28.064*
	ToPO	**21.858**	**0.2894**	**0.829**	**5.361**	**28.081**
SDXL	Base	22.375	0.2790	0.816	5.343	28.062
	Diffusion-DPO	*22.635*	0.2938	**1.096**	*5.358*	**28.655**
	SFT-winners	21.974	0.2772	0.787	5.338	28.529
	InPO	**22.657**	*0.2953*	1.027	5.350	28.498
	MaPO	22.410	0.2868	0.935	**5.364**	28.190
	ToPO	**22.657**	**0.2954**	*1.064*	5.357	*28.604*

attributor and should therefore be interpreted as a diagnostic of the attribution rule rather than a direct measure of localization in the final trained model.

Table 8: Expanded T2I-CompBench results.

Backbone	Method	CB Avg.† ↑	Color ↑	Shape ↑	Texture ↑	Count ↑	Spatial ↑	Non-spatial ↑
SD-1.5	Base	0.3393	0.3750	0.3715	0.4201	0.4514	0.1068	0.3109
	Diffusion-DPO	0.3521	0.4028	0.3842	0.4243	0.4591	0.1309	0.3114
	SFT-winners	0.3870	0.4761	0.4310	0.4773	0.4677	0.1572	0.3129
	KTO	0.3887	0.4875	0.4300	0.4729	0.4748	0.1538	0.3131
	InPO	*0.3984*	*0.4967*	*0.4389*	*0.4966*	*0.4785*	*0.1663*	*0.3135*
	ToPO	**0.4123**	**0.5242**	**0.4579**	**0.5175**	**0.4793**	**0.1816**	**0.3136**
SDXL	Base	0.4533	0.6277	0.4878	0.5587	0.5114	0.2179	0.3161
	Diffusion-DPO	*0.4974*	*0.7177*	*0.5378*	*0.6436*	**0.5402**	**0.2262**	0.3190
	SFT-winners	0.4436	0.6089	0.4955	0.5437	0.5046	0.1888	**0.3202**
	InPO	0.4731	0.6731	0.5231	0.5902	0.5176	0.2170	0.3174
	MaPO	0.4691	0.6429	0.5197	0.5808	0.5302	0.2235	0.3175
	ToPO	**0.4975**	**0.7190**	**0.5398**	**0.6441**	*0.5369*	*0.2250*	*0.3201*

Table 9: Held-out HPDv2 results on 3,200 prompts using one random training seed.

Backbone	Method	PickScore ↑	HPSv2 ↑	ImageReward ↑	Aesthetic ↑	CLIP ↑
SD-1.5	Base	20.825	0.2416	0.105	5.327	29.343
	Diffusion-DPO	21.272	0.2551	0.345	5.339	29.703
	SFT-winners	21.605	0.2865	0.770	5.357	30.253
	KTO	21.526	0.2842	0.725	5.355	29.947
	InPO	*21.854*	*0.2907*	*0.847*	*5.362*	**30.600**
	ToPO	**21.901**	**0.2957**	**0.907**	**5.364**	*30.515*
SDXL	Base	22.547	0.2881	0.934	5.356	30.281
	Diffusion-DPO	22.884	0.3049	*1.122*	*5.363*	*30.734*
	SFT-winners	22.040	0.2877	0.872	5.347	30.434
	InPO	*22.928*	**0.3113**	1.095	5.361	30.637
	MaPO	22.617	0.2988	1.023	**5.371**	30.464
	ToPO	**22.941**	*0.3063*	**1.123**	5.360	**30.821**

Table 10: Token-modulation diagnostic with the frozen SD-1.5 reference attributor (60 cases per class).

Class	Mean R	Ratio of means	95% CI	Raw p	Holm p	$R > 1/R < 1$
Color	1.404	1.402	[1.286, 1.523]	3.4×10^{-7}	1.7×10^{-6}	48/12
Shape	1.261	1.256	[1.136, 1.390]	1.1×10^{-3}	5.7×10^{-3}	39/21
Texture	1.310	1.290	[1.205, 1.419]	8.9×10^{-6}	4.4×10^{-5}	43/17
Count	.941	.936	[.851, 1.032]	.140	.140	23/37
Spatial	.918	.906	[.817, 1.024]	.140	.140	24/36

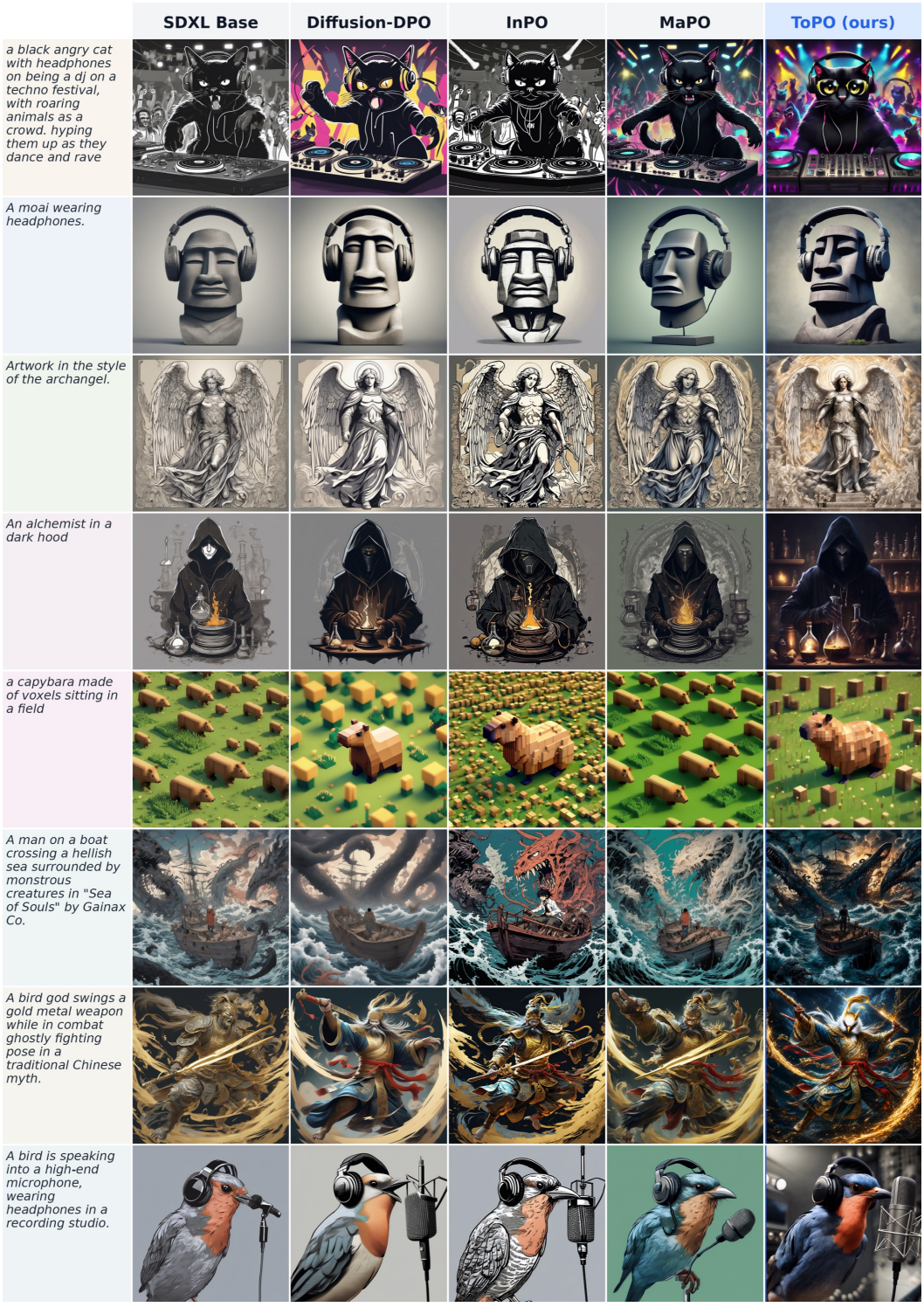

Figure 6: Additional fixed-prompt SDXL aesthetic comparisons. Columns compare SDXL Base, Diffusion-DPO, InPO, MaPO, and ToPO on stylized subjects, visual detail, material rendering, and complex scenes.

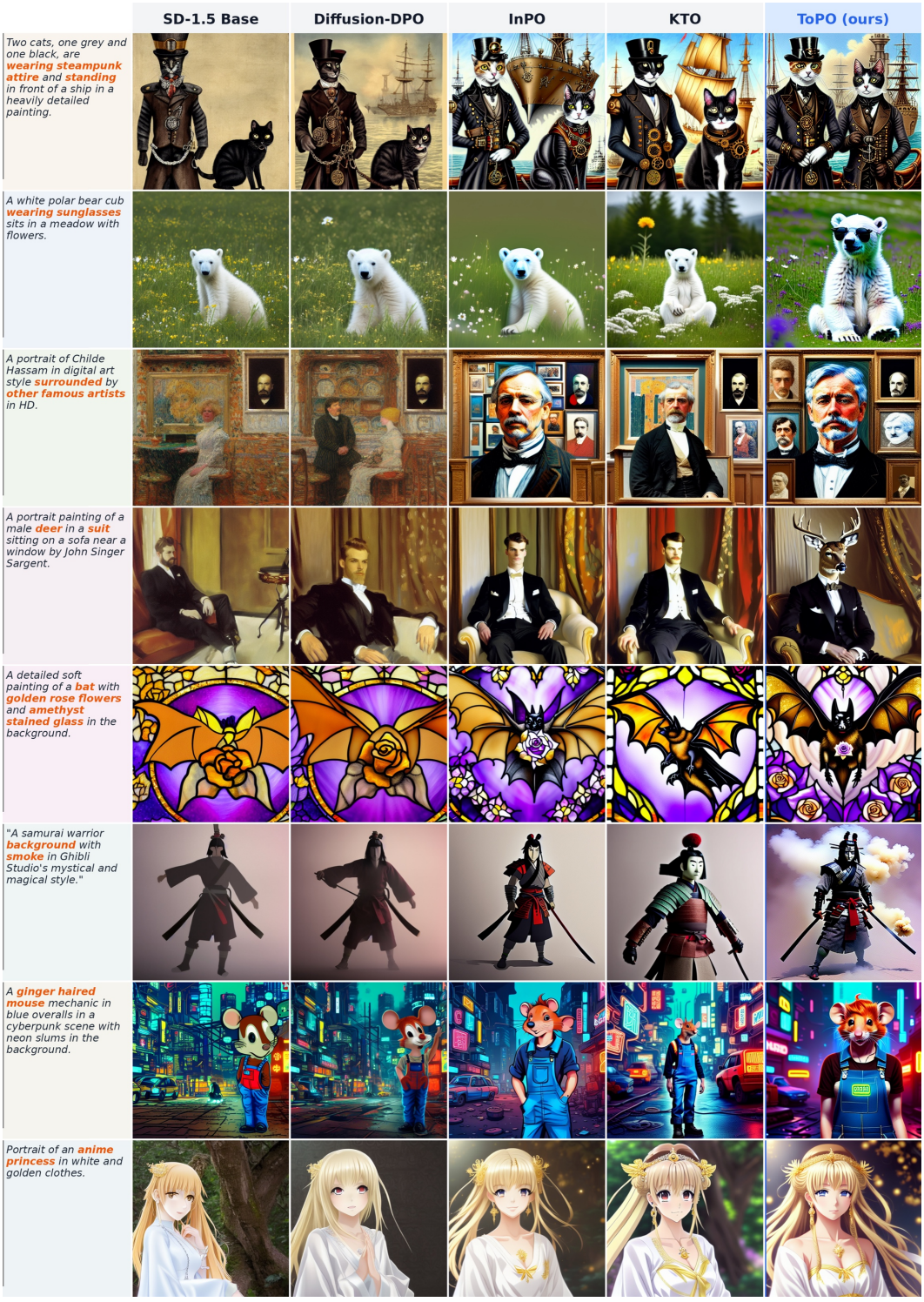

Figure 7: Additional fixed-prompt SD-1.5 visual-quality comparisons. Columns compare SD-1.5 Base, Diffusion-DPO, InPO, KTO, and ToPO on eight prompts with detailed apparel, accessories, materials, stylized subjects, and scenes.

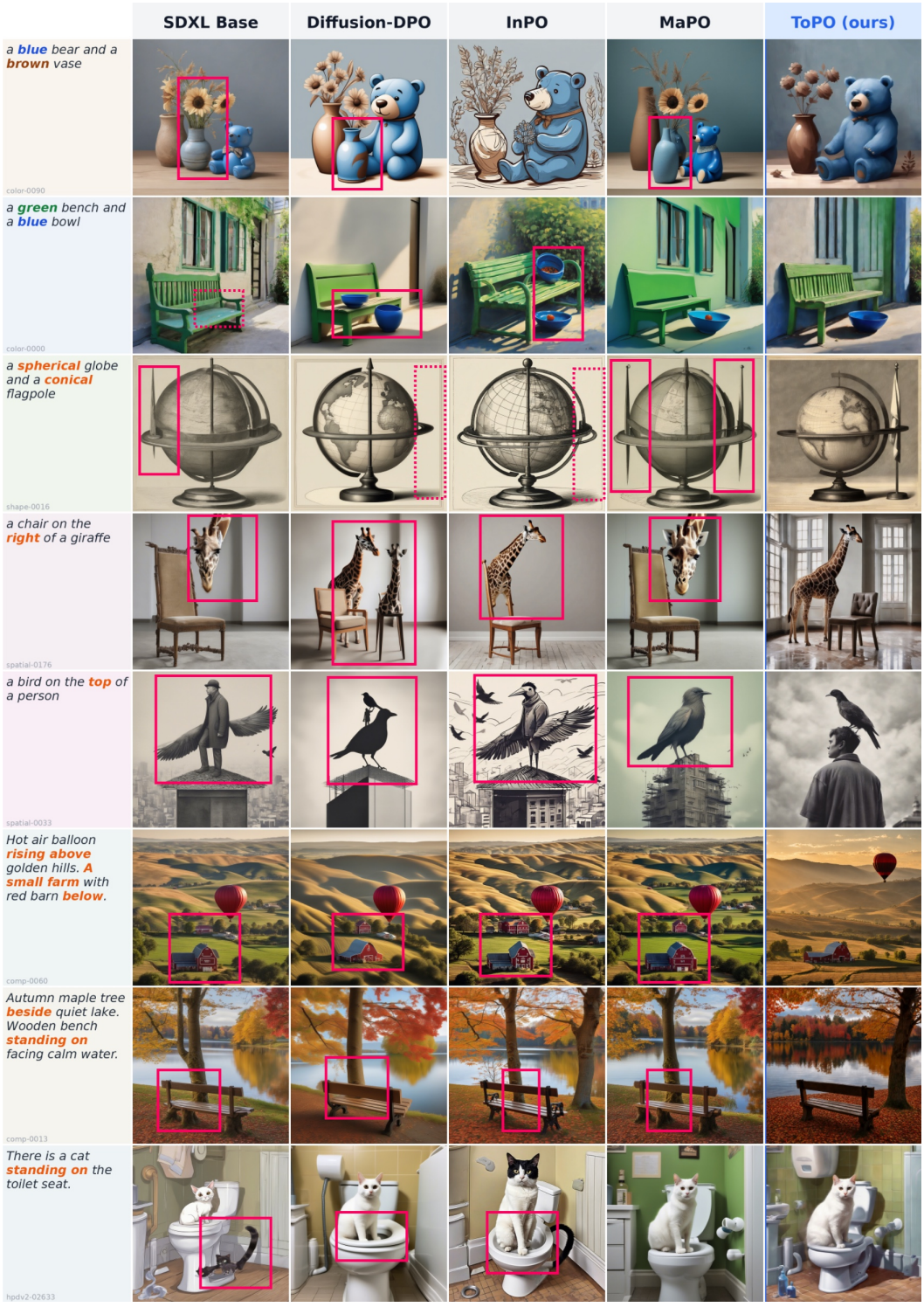

Figure 8: Additional fixed-prompt SDXL attribute and relational comparisons. Highlighted words denote requested colors, shapes, or relations. Columns compare SDXL Base, Diffusion-DPO, InPO, MaPO, and ToPO.



Figure 9: **Seed-wise samples from the final SDXL ToPO checkpoint.** Each row fixes one prompt, while the five image columns use different random seeds. The plate illustrates variation in composition, viewpoint, and local detail across seeds.



Figure 10: **Additional visual-quality samples from the final SDXL ToPO checkpoint.** The gallery covers detailed environments, atmospheric scenes, fantasy subjects, portraits, and material-rich natural elements.

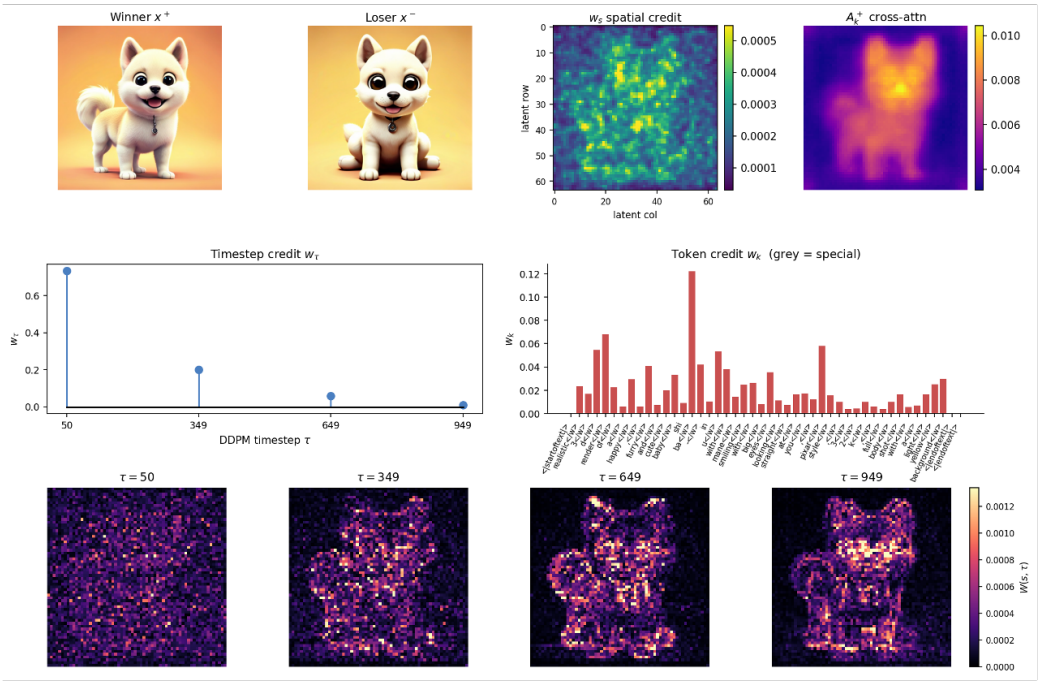

Figure 11: Token-conditioned spatial–temporal routing in ToPO. The held-out Pick-a-Pic v2 winner–loser pair is conditioned on the prompt "Realistic 3d render of a happy, furry and cute baby Shiba-Inu with mane smiling with big eyes looking straight at you, Pixar style, 32k, full body shot with a light yellow background." The figure shows the spatial factor w_s, timestep factor w_τ, content-token modulation w_k, and composed field $W(s, \tau)$ at four denoising anchors. This final-checkpoint diagnostic is distinct from the frozen-reference route used during training.