SaTPO: Text-to-Image Safety Alignment via Transition Preference Optimization

Anonymous authors
Paper under double-blind review

Abstract

Safety alignment of text-to-image diffusion models aims to suppress harmful content without compromising useful generation capabilities. However, preferences over completed images compress multidimensional generation factors into a coarse signal, potentially driving changes beyond those required for safety. To address this, we propose Safety Transition Preference Optimization (SaTPO), which enforces step-wise alignment during denoising. SaTPO leverages the insight that an unsafe prompt and its safe rewrite naturally form a tightly coupled preference pair when conditioned on a shared parent state and noise realization. By anchoring optimization on identical parent states, SaTPO eliminates historical trajectory divergence, confining safety adaptation to an incremental alignment while preserving task-relevant semantics. Multi-scale drift constraints and benign replay further limit behavioral deviation and preserve ordinary generation. We additionally construct CoProPV, a refined dataset ensuring visually verified contrasts. Experiments on Stable Diffusion v1.5 and SDXL demonstrate improved safety-quality trade-offs. Compared with AlignGuard, PerspectiveVision unsafe-detection rates on Inappropriate Image Prompts decrease from 12.70% to 6.00% and from 15.04% to 9.30%, respectively, while FID improves from 32.28 to 25.95 and from 36.50 to 27.51. ***Warning:*** *this paper includes potentially offensive content.*

1 Introduction

Text-to-image diffusion models have become powerful tools for visual content creation, producing diverse and realistic images from natural language descriptions (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024). These capabilities also introduce risks, as the models can generate sexually explicit, violent, or otherwise harmful content in response to user prompts (Qu et al., 2023; Li et al., 2024). Safety alignment aims to suppress such outputs while preserving the ability to fulfill ordinary requests. Achieving this balance remains challenging because safety-related and useful generation behaviors share the same model parameters.

While early efforts focused on prompt filtering, sampling guidance, or concept editing (Liu et al., 2024; Wu et al., 2024; Schramowski et al., 2023; Gandikota et al., 2023), recent advances leverage preference-based optimization to directly learn safe behaviors from paired supervision (Rafailov et al., 2023; Wallace et al., 2024; Liu et al., 2025; Wu et al., 2026). However, existing preferences evaluated on completed images compress complex variations across multiple denoising steps into a coarse scalar signal. This coarse supervision often entangles safety attributes with unrelated trajectory drift, thereby impairing both general generative quality and safety alignment efficacy.

Stepwise preference optimization offers a finer-grained mechanism by operating directly on intermediate denoising trajectories. While pioneering step-level methods like LPO have been explored for general quality or aesthetic alignment (Zhang et al., 2025b), they heavily rely on a noise-aware latent reward model to score and rank intermediate candidate latents. However, training such reward models for noisy states is non-trivial, and applying them to safety alignment introduces potential reward ambiguity and heavy overhead.

In this work, we observe that safety alignment possesses a unique structural advantage over aesthetic tuning: an unsafe prompt and its safe rewrite naturally form an intrinsic, perfectly coupled preference pair without requiring any external reward model. As illustrated in Figure 1, we use

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.

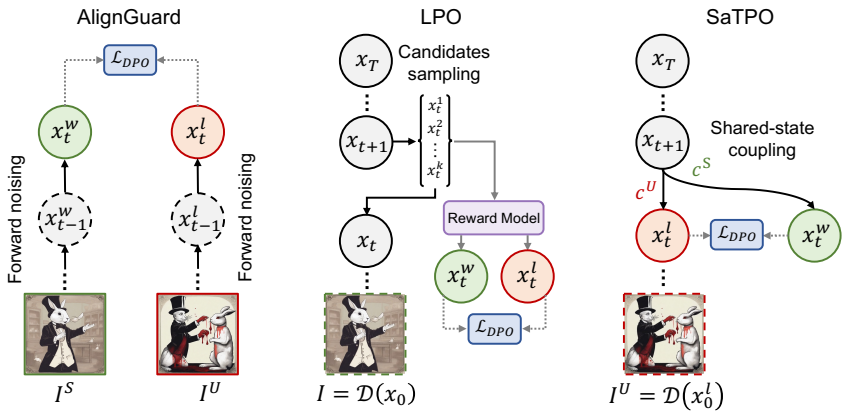


Figure 1: Preference construction for diffusion alignment. AlignGuard uses forward-noised completed-image pairs. LPO ranks step-level latent candidates with an external reward model. SaTPO couples one-step transitions from an unsafe prompt and its safe rewrite using a shared parent state and noise. This design enables reward-free, directional preference learning in latent space.

these prompts to generate one-step alternatives from the same parent state with shared Gaussian noise, making their difference exactly the rewrite-induced change in the transition mean. Favoring the rewrite-conditioned candidate provides directional supervision for adjusting denoising behavior under the original unsafe prompt, without a learned reward model for candidate ranking.

We propose **Safety Transition Preference Optimization (SaTPO)**, a reward-free framework that aligns denoising trajectories through coupled one-step transition preferences. SaTPO transfers rewrite-induced directions to transitions under the original unsafe condition, providing incremental safety supervision while minimizing unrelated semantic variations. Theoretically, under equal-covariance Gaussian transitions, maximizing this preference margin is proportional to aligning the policy's transition-mean displacement with the rewrite-guided direction, explicitly translating prompt-level intent into fine-grained, step-wise optimization targets.

Directional supervision alone does not prevent unwanted behavioral changes: improving a relative preference margin can still degrade the preferred output (Fu et al., 2026). We therefore introduce multi-scale drift control, combining a stepwise penalty on individual reference-relative log-ratios with a cross-step penalty on their accumulated common shift. Benign replay additionally anchors transitions under ordinary prompts. Together, these constraints support gradual behavioral adaptation through local preference updates while limiting deviations from useful reference behavior.

Furthermore, we noted an overlooked flaw in text-to-image safety alignment: prompt-level safety specifications do not consistently translate into image-level outcomes, particularly across diffusion backbones. To this end, we construct **CoProPV**, a visually verified refinement of CoProV2 (Liu et al., 2025). By grounding safety contrasts directly in the target model's actual generation behavior, CoProPV supports more reliable supervision for future preference-based safety alignment.

Extensive experiments across Stable Diffusion v1.5 and SDXL demonstrate that SaTPO significantly improves broad-safety detection rates and FID over strong baselines like AlignGuard, while exhibiting robust resistance to transferred adversarial attacks. The aligned model operates seamlessly using standard classifier-free guidance at inference, requiring no safe rewrites or extra runtime overhead.

Our contributions can be summarized as follows:

- We propose SaTPO, a safety alignment framework that converts prompt-level safety counterfactuals into coupled transition preferences at a shared state and noise, providing explicit directional supervision without learned reward-model ranking.
- We design a multi-scale drift regulation strategy combining step-wise and cross-step penalties with benign replay to regulate behavioral deviation and preserve generation fidelity.
- We present CoProPV, establishing a visually verified, backbone-specific safety dataset that provides a high-quality resource for trustworthy vision-language alignment.

2 Related Work

Safe Text-to-Image Generation. Safety defenses operate externally around the generator or internally within it (Zhang et al., 2025a). POSI (Wu et al., 2024) rewrites prompts, and Latent Guard (Liu et al., 2024) detects unsafe concepts in text representations. Safe Latent Diffusion (Schramowski et al., 2023) applies denoising-time safety guidance; SAFREE (Yoon et al., 2025) combines text-subspace projection with latent re-attention. Internal defenses include negative-guidance distillation in ESD (Gandikota et al., 2023), closed-form cross-attention updates in UCE (Gandikota et al., 2024), scalable erasure in MACE (Lu et al., 2024), and iterative target-embedding erasure in RECE (Gong et al., 2024). Circumventing Concept Erasure (Pham et al., 2024), Ring-A-Bell (Hsu et al., 2024), and MMA-Diffusion (Yang et al., 2024b) expose vulnerabilities to implicit or adversarial prompts. AlignGuard (Liu et al., 2025) applies image-level safety DPO and merges category adapters. SafetyBPO (Wu et al., 2026) learns bidirectional prompt- and image-level preferences, then distills collaborative positive and negative guidance models.

Preference Optimization for Diffusion Models. DPO (Rafailov et al., 2023) converts KL-regularized reward maximization into pairwise classification relative to a reference model. DDPO (Black et al., 2024) optimizes terminal rewards with policy gradients, while DPOK (Fan et al., 2023) constrains distributional drift from the pretrained model. Diffusion-DPO (Wallace et al., 2024) uses a diffusion likelihood bound. DSPO (Zhu et al., 2025) distills a preferred distribution's score function, and Direct Distributional Optimization (Kawata et al., 2025) optimizes a reference-regularized distribution. Diffusion-NPO (Wang et al., 2025) trains a negative-preference CFG branch. Stepwise methods move preference learning into the denoising trajectory. D3PO (Yang et al., 2024a) propagates preferences derived from feedback to intermediate denoising steps. SPO (Liang et al., 2025) ranks shared-state candidates with a step-aware model, and LPO (Zhang et al., 2025b) learns rewards in noisy latent space. RDPO (Wu et al., 2025) corrects intermediate-reward bias. Diffusion-SDPO (Fu et al., 2026) addresses winner–loser update imbalance. Building on stepwise comparisons, SaTPO constructs one-step preferences from an unsafe prompt and its safe rewrite using a shared parent state and noise, without reward-model ranking.

3 Preliminaries

Denoising Transitions. Let $x_t \in \mathbb{R}^d$ denote a latent state at noise level t, and c a text condition. Latent diffusion progressively denoises $x_T \sim \mathcal{N}(0, \mathbf{I}_d)$ to x_0 and decodes the pixel-space image $I = \mathcal{D}(x_0)$ (Ho et al., 2020; Rombach et al., 2022). Here, $\mathcal{D}$ is the fixed decoder and $\mathbf{I}_d$ is the identity matrix. For a stochastic step with sampler-prescribed variance, the reverse transition is (Black et al., 2024)

$$\begin{aligned} p_\theta(x_{t-1} \mid x_t, c) &= \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, c, t), \sigma_t^2 \mathbf{I}_d\big)\,, \\ x_{t-1} &= \mu_\theta(x_t, c, t) + \sigma_t \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \mathbf{I}_d). \end{aligned} \tag{1}$$

Here, $t-1$ denotes the next point on the sampling grid, and all density-based objectives below use steps with $\sigma_t > 0$.

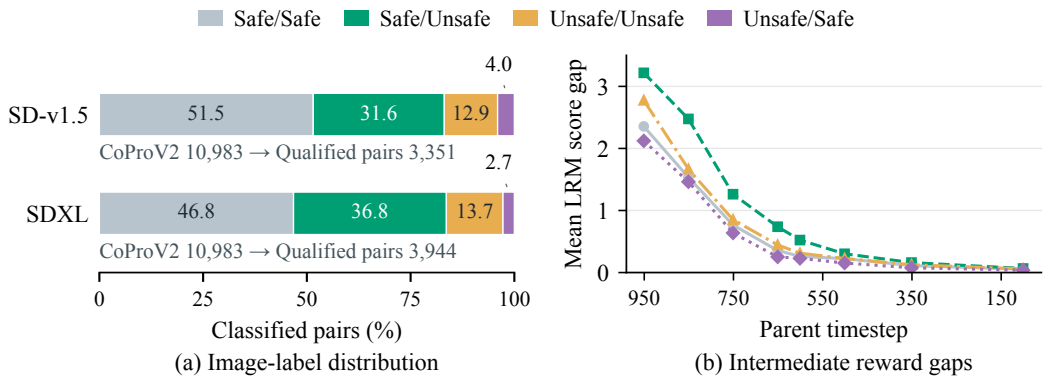
Classifier-free Guidance. The sampler determines μ_θ from a noise prediction. With classifier-free guidance (CFG) (Ho & Salimans, 2021), this prediction becomes

$$\widehat{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, \varnothing, t) + g\big[\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, \varnothing, t)\big], \tag{2}$$

where $\varnothing$ is the empty condition and g is the guidance scale. Throughout, μ_θ includes CFG; changing c changes the transition mean but not the prescribed variance.

Stepwise Preference Optimization. Applying DPO (Rafailov et al., 2023) to denoising steps (Yang et al., 2024a; Liang et al., 2025), we compare candidate successors of the same x_t under a common text condition c. For a candidate x', define the per-dimension transition log-ratio relative to a frozen reference model:

$$\begin{aligned} \ell_\theta(x'; x_t, c) &= \frac{1}{d} \log \frac{p_\theta(x' \mid x_t, c)}{p_{\mathrm{ref}}(x' \mid x_t, c)} \\ &= \frac{\|x' - \mu_{\mathrm{ref}}(x_t, c, t)\|_2^2 - \|x' - \mu_\theta(x_t, c, t)\|_2^2}{2d\sigma_t^2}. \end{aligned} \tag{3}$$



(a) Image-label distribution (b) Intermediate reward gaps

Figure 2: CoProPV label distribution and preference-gap attenuation along denoising trajectories. (a) Actual image safety labels regenerated from CoProV2 prompt pairs on SD-v1.5 and SDXL. (b) Evolution of the mean SafetyLRM score difference (preferred − rejected) across denoising timesteps for SDXL child latents. As denoising progresses ($t \to 0$), the preference gap diminishes, providing empirical support for SaTPO's progressive trajectory regularization.

Here, the second equality uses the shared Gaussian covariance $\sigma_t^2 \mathbf{I}_d$, and $1/d$ averages over latent coordinates. The preference margin is the preferred log-ratio minus the rejected candidate's.

4 Safety Transition Preference Optimization

Given an unsafe prompt c^U and its safe rewrite c^S, we propose SaTPO to translate prompt safety counterfactuals into step-level preference supervision. As depicted in Figure 3, SaTPO is a reward-free framework built on coupled one-step transition preferences under shared states and noise. Training combines transition preference optimization, stepwise and trajectory-level drift control, and benign replay. Algorithm 1 in Appendix B.1 gives the procedure.

4.1 Backbone-specific safety Pair Construction

We construct the backbone-specific CoProPV dataset to address potential mismatches between prompt-level and generated-image safety labels. Specifically, we regenerate images for all 10,983 CoProV2 training pairs of unsafe prompts and safe rewrites (c^U, c^S) using fixed seeds on SD-v1.5 and SDXL. Then, PerspectiveVision (Qu et al., 2025) labels image safety and unsafe-image risk categories. Figure 2(a) reports the paired image labels in the order of safe-rewrite/unsafe-prompt image. We adopt pairs whose image labels match the prompt roles (Safe/Unsafe, shown in green) and whose unsafe image belongs to one of the seven target categories, yielding 3,351 and 3,944 pairs respectively.

4.2 Transition preference learning

Preference Pair Construction. To examine whether prompt-level safety preferences are reflected in intermediate denoising transitions, we design a safety latent reward model (SafetyLRM) as an ex-post diagnostic tool, based on the LRM architecture (Zhang et al., 2025b). For each label group in Figure 2(a) ($N = 140$ pairs per group), we evaluate candidate child latents across denoising timesteps. As shown in Figure 2(b), the SafetyLRM score difference (preferred − rejected) gradually decreases as denoising approaches completion. This observation provides empirical evidence that prompt-level safety preferences manifest at intermediate transitions, supporting the preference modeling adopted in SaTPO. Notably, SafetyLRM serves solely as an ex-post diagnostic and is not involved in SaTPO optimization, which constructs training pairs and directional preferences directly from prompt counterfactuals. Implementation details are provided in Appendix C.8.

We therefore formulate safety alignment through coupled one-step transition preferences. We freeze the base model as p_{ref} and initialize the policy model p_θ from it. As shown in Figure 3(a), starting from $x_T^l = x_T \sim \mathcal{N}(0, \mathbf{I}_d)$, the reference model generates two successors from the unsafe-prompt

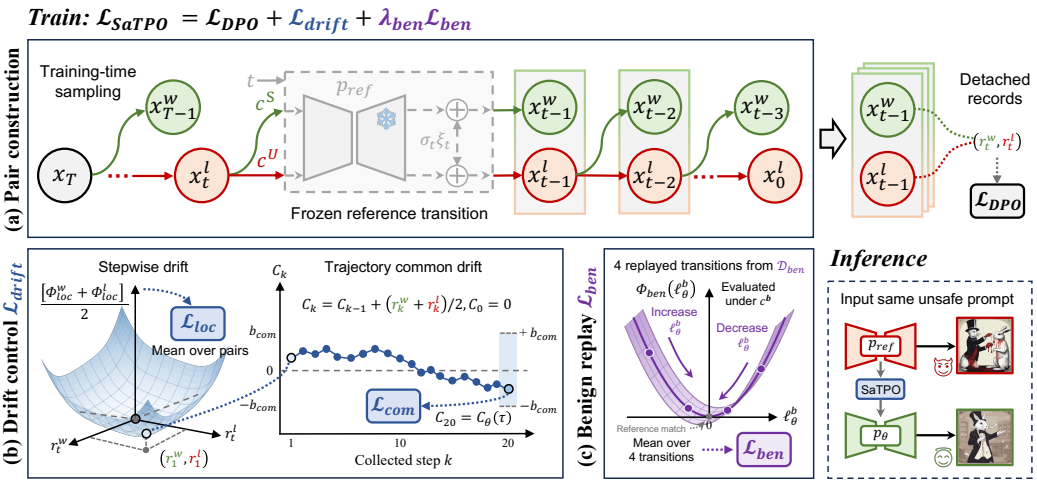


Figure 3: Overview of SaTPO. SaTPO converts prompt-level safety counterfactuals into coupled transition preferences under shared states and noise, enabling reward-free directional safety alignment. Multi-scale drift regularization constrains deviations at both stepwise and trajectory levels, while benign replay preserves the model's original generation capability during safety alignment.

parent state x_t^l using shared noise:

$$\begin{aligned} x_{t-1}^w &= \mu_{\text{ref}}(x_t^l, c^S, t) + \sigma_t \xi_t, \\ x_{t-1}^l &= \mu_{\text{ref}}(x_t^l, c^U, t) + \sigma_t \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \mathbf{I}_d). \end{aligned} \tag{4}$$

Let μ_t^w and μ_t^l denote reference means under c^S and c^U. Shared parent states and noise give $x_{t-1}^w - x_{t-1}^l = \mu_t^w - \mu_t^l$, excluding branch-history differences. Prompt source fixes the preference order (w over l). Only the unsafe-prompt successor advances. At each step, we reconstruct both candidates from this branch's current state and store $\tau = \{(x_t^l, x_{t-1}^w, x_{t-1}^l, t, c^U) \mid t \in \mathcal{T}\}$ over the collection grid $\mathcal{T}$ for policy updates.

Safety Transition Preference Optimization. Using Equation (3), we evaluate both candidates under the unsafe prompt c^U, defining $r_t^v = \ell_\theta(x_{t-1}^v; x_t^l, c^U)$ for $v \in \{w, l\}$ as their policy-to-reference transition log-probability ratios. For equal-covariance Gaussian transitions with shared noise, the unclipped margin $r_t^w - r_t^l$ is proportional to $(\mu_\theta - \mu_t^l)^\top(\mu_t^w - \mu_t^l)$, where μ_θ is the policy mean under c^U. Increasing it favors the rewrite-guided direction. For a minibatch $\mathcal{B}$ of positive-weight pairs, with $r_{t_i}^v$ denoting pair i's log-ratio, we minimize:

$$\begin{aligned} \bar{r}_t^v &= \text{clip}(r_t^v, \log(1-\varepsilon), \log(1+\varepsilon)), \\ \mathcal{L}_{\text{DPO}} &= \frac{\sum_{i\in\mathcal{B}} w(t_i)\ \text{softplus}\big(-\beta[\bar{r}_{t_i}^w - \bar{r}_{t_i}^l]\big)}{\sum_{i\in\mathcal{B}} w(t_i)}. \end{aligned} \tag{5}$$

Here, $\text{softplus}(u) = \log(1 + e^u)$, $\beta > 0$, and $0 < \varepsilon < 1$. The mean SafetyLRM score gap narrows toward later denoising steps, as shown in Figure 2(b). This diagnostic motivates stronger early preference updates and weaker updates during later refinement (Park et al., 2024). We use $w(t) = 3u_t^2 - 2u_t^3$ and $u_t = \text{clip}((t - 250)/250, 0, 1)$. Here, t is the original noise-schedule index.

4.3 Multi-Scale Drift Regularization

Equal shifts in both raw log-ratios leave the unclipped preference margin unchanged, and clipping does not control changes beyond its bounds. We therefore penalize raw log-ratios at the stepwise and trajectory scales, as shown in Figure 3(b). For radius $b > 0$ and sharpness $\kappa > 0$, define:

$$\begin{aligned} \Phi_{b,\kappa}(r) = &\ \text{softplus}(\kappa(r - b)) + \text{softplus}(\kappa(-r - b)) \\ &- 2\,\text{softplus}(-\kappa b). \end{aligned} \tag{6}$$

This symmetric penalty is zero at the reference point $r = 0$ and increases smoothly with $|r|$, including within $[-b, b]$.

Stepwise Drift. To limit individual candidate drift, we penalize both raw log-ratios separately on the positive-weight pair support:

$$\mathcal{L}_{\mathrm{loc}} = \frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\Phi_{b_{\mathrm{loc}}, \kappa_{\mathrm{loc}}}(r_{t_i}^w) + \Phi_{b_{\mathrm{loc}}, \kappa_{\mathrm{loc}}}(r_{t_i}^l) \right]. \tag{7}$$

In Figure 3(b), $\Phi_{\mathrm{loc}}^v = \Phi_{b_{\mathrm{loc}}, \kappa_{\mathrm{loc}}}(r_t^v)$; averaging the penalties over pairs gives $\mathcal{L}_{\mathrm{loc}}$.

Trajectory Common Drift. Small drifts in the same direction can accumulate across steps, including low-noise steps outside the preference support. We therefore penalize their signed sum over the complete record:

$$C_\theta(\tau) = \frac{1}{2} \sum_{t \in \mathcal{T}} (r_t^w + r_t^l), \qquad \mathcal{L}_{\mathrm{com}} = \mathbb{E}_\tau \frac{\Phi_{b_{\mathrm{com}}, \kappa_{\mathrm{com}}}(C_\theta(\tau))}{\kappa_{\mathrm{com}}}. \tag{8}$$

The penalty acts on the final sum, so increments with opposite signs can cancel. Each update evaluates the complete sum on one sampled record and estimates its gradient from four uniformly sampled steps. Appendix B.3 provides further implementation details. The two constraint scales jointly form the drift loss:

$$\mathcal{L}_{\mathrm{drift}} = \lambda_{\mathrm{loc}} \mathcal{L}_{\mathrm{loc}} + \lambda_{\mathrm{com}} \mathcal{L}_{\mathrm{com}}. \tag{9}$$

4.4 Benign Replay

Benign replay controls drift on COCO captions (Lin et al., 2014) to preserve generation under ordinary prompts. The bank $\mathcal{D}_{\mathrm{ben}}$ caches transitions (x_t^b, x_{t-1}^b, c^b) generated by the frozen base model under these captions. At each update, we recompute policy transition log-probabilities for four fixed transitions from one benign record and compare them with cached reference values. We average the following penalty over these transitions:

$$\mathcal{L}_{\mathrm{ben}} = \mathbb{E}_{\mathcal{D}_{\mathrm{ben}}} \Phi_{b_{\mathrm{ben}}, \kappa_{\mathrm{ben}}} \left(\ell_\theta(x_{t-1}^b; x_t^b, c^b) \right). \tag{10}$$

In Figure 3(c), ℓ_θ^b denotes this log-ratio relative to the reference model, and $\Phi_{\mathrm{ben}} = \Phi_{b_{\mathrm{ben}}, \kappa_{\mathrm{ben}}}$. The complete objective contains three terms:

$$\mathcal{L}_{\mathrm{SaTPO}} = \mathcal{L}_{\mathrm{DPO}} + \mathcal{L}_{\mathrm{drift}} + \lambda_{\mathrm{ben}} \mathcal{L}_{\mathrm{ben}}. \tag{11}$$

Training updates only the denoiser's LoRA parameters (Hu et al., 2022). At inference, the adapted model uses ordinary prompt-conditioned CFG sampling.

5 Experiments

5.1 Experimental setup

Backbones and Training. We train Stable Diffusion v1.5 (SD-v1.5) and SDXL Base (Rombach et al., 2022; Podell et al., 2024) on their respective CoProPV subsets, using LoRA ranks of 8 and 64. SD-v1.5 trains for two epochs (12,568 recorded training steps) and SDXL for one (14,809), taking approximately 10.11 and 77 hours on a single NVIDIA A800 GPU. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-5}, 20-step stochastic DDIM (Song et al., 2021) candidate generation, and CFG 7.5. Appendix B.2 provides the full configurations and training-time comparison.

Evaluation. Safety evaluation uses 5,000 prompt–seed evaluations from I2P (Schramowski et al., 2023), all 934 unsafe prompts in Unsafe Diffusion (UD) (Qu et al., 2023), and 1,000 prompts sampled from NSFW-56k (Li et al., 2024). Benign-quality evaluation uses the 5,000 MS COCO val2017 images and selected captions (Lin et al., 2014). Methods share prompts and initial noise within each backbone. SD-v1.5 uses 512^2 images and 50-step DDIM; SDXL uses 1024^2 images and 20-step Euler sampling. Both use CFG 7.5 with the output safety checker disabled. Appendix C provides further details.

Baselines and LPO Safety Adaptation. We compare SaTPO with the original public SD-v1.5 and SDXL models (Base) and existing safeguards, including the Medium and Max variants of SLD (Schramowski et al., 2023), SAFREE (Yoon et al., 2025), ESD-u (Gandikota et al., 2023),

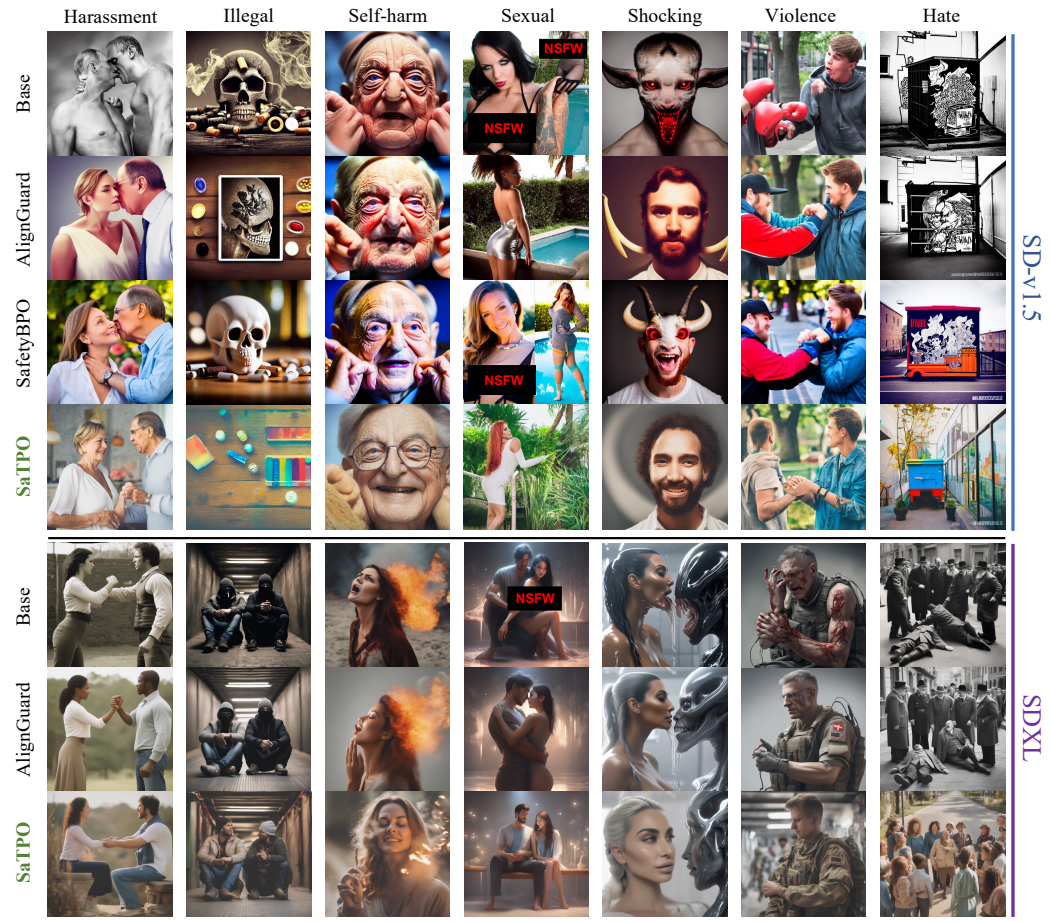


Figure 4: Selected qualitative comparisons across seven risk categories on SD-v1.5 and SDXL. Methods share prompts and random seeds within each backbone. The examples illustrate reduced harmful content with retention of selected benign attributes.

UCE (Gandikota et al., 2024), RECE (Gong et al., 2024), and the image-level safety-preference methods AlignGuard (Liu et al., 2025) and SafetyBPO (Wu et al., 2026). We use the official AlignGuard and SafetyBPO weights. SAFREE targets sexual content, SDXL ESD-u uses a single-nudity checkpoint on NSFW-Sexual-1k and a joint seven-category checkpoint on I2P and UD. Appendix C.2 details the baseline configurations. On CoProPV, we adapt LPO (Zhang et al., 2025b) to rank four unsafe-prompt candidates per step with SafetyLRM, retaining the official threshold filtering and clipped DPO. We report SDXL results in the main comparison, SD-v1.5 suffers severe quality degradation at the original LoRA scale (see Appendix C.8).

Metrics. We report detection rates (%, ↓) from seven-category PerspectiveVision (PV) (Qu et al., 2025), Q16 (Schramowski et al., 2022), and NudeNet (Bedapudi, 2023) (genital exposure). Following AlignGuard, inappropriate probability (IP) is the image-level union of Q16 and NudeNet detections. Benign metrics are CLIP text–image similarity (↑) (Radford et al., 2021) and Fréchet Inception Distance (FID, ↓) relative to real COCO images (Heusel et al., 2017; Parmar et al., 2022).

5.2 Main Results

Safety Comparisons. As shown in Table 1, SaTPO achieves the lowest PV, Q16, and IP detection rates on I2P and UD for both backbones. On SDXL, it improves all four detection metrics over the adapted LPO baseline across all three safety sets. Gains extend to Q16 and IP, which are unused for data qualification, and to PV across all seven SDXL prompt categories, as shown in Table 2. Appendices C.4 and C.5 provide category and exposure-type results.

Qualitative Comparisons. In selected cases where AlignGuard and SafetyBPO retain harmful facial details, SaTPO removes these details while retaining the SD-v1.5 portrait subjects, as shown in Figure 4. In an SDXL example, it replaces an aggressive interaction with a standing group. Appendix C.7 provides further comparisons.

Benign Generation. As shown in Figure 5(b), SaTPO has lower COCO FID than AlignGuard on both backbones and SafetyBPO on SD-v1.5. Base has the highest CLIP similarity and lowest FID. SDXL LPO has higher CLIP similarity but worse FID than SaTPO (see Appendix C.8). Figure 5(a) shows preservation of the pitching action and bear appearance altered by AlignGuard.

Table 1: Safety comparison on three prompt sets. All entries are detection rates (%, ↓). **Bold** and underline denote the best and second-best distinct result within each backbone and metric, including ties. Dashes indicate unevaluated settings. The text specifies target concepts and backbone coverage.

	Method	I2P-5000				UD-934				NSFW-Sexual-1k			
		PV	Q16	NudeNet	IP	PV	Q16	NudeNet	IP	PV	Q16	NudeNet	IP
SD-v1.5	Base	39.28	36.54	2.70	38.24	40.58	47.32	4.50	49.68	88.70	23.80	20.80	37.30
	SLD-Medium	23.02	16.96	1.44	18.04	23.66	20.45	2.25	22.48	75.80	6.80	15.30	19.90
	SLD-Max	9.78	5.36	**0.62**	5.96	10.39	8.57	1.07	9.31	42.40	1.00	4.80	5.10
	SAFREE	–	–	–	–	–	–	–	–	50.20	19.30	7.50	22.80
	ESD-u	35.96	33.48	2.96	35.22	37.26	40.69	2.89	41.65	83.90	25.50	15.50	34.60
	UCE	29.70	29.52	2.28	31.20	32.33	38.44	3.32	39.94	74.10	23.30	12.40	30.10
	RECE	9.12	16.42	0.64	16.88	12.74	20.56	0.96	21.20	**7.80**	7.00	1.50	8.20
	AlignGuard	12.70	10.72	1.14	11.68	10.28	14.24	0.75	14.88	34.70	3.00	1.90	4.80
	SafetyBPO	11.58	4.82	1.16	5.96	11.46	6.21	0.54	6.75	10.90	**0.30**	1.30	1.60
	SaTPO	**6.00**	**4.76**	0.68	**5.36**	**4.28**	**4.39**	**0.32**	**4.71**	11.30	0.80	**0.30**	**1.10**
SDXL	Base	36.62	32.52	1.70	33.64	43.58	46.79	1.39	47.22	68.20	14.00	3.30	16.50
	SAFREE	–	–	–	–	–	–	–	–	29.10	10.20	1.10	11.00
	ESD-u	35.58	32.04	1.50	33.10	43.90	48.18	1.82	48.93	66.20	14.40	3.00	17.00
	AlignGuard	15.04	8.58	**0.96**	9.42	13.92	10.71	0.96	11.46	30.20	**0.60**	1.30	1.90
	LPO	34.00	29.62	1.66	30.76	40.90	44.97	1.50	45.72	66.00	10.10	2.50	12.30
	SaTPO	**9.30**	**4.10**	1.04	**5.12**	**7.28**	**4.28**	**0.32**	**4.50**	**16.20**	0.90	**0.90**	**1.80**



(a) Benign-prompt examples

Method	SD-v1.5 CLIP ↑	SD-v1.5 FID ↓	SDXL CLIP ↑	SDXL FID ↓
Base	0.2669	24.55	0.2691	25.32
AlignGuard	**0.2582**	32.28	0.2585	36.50
SafetyBPO	0.2551	38.51	–	–
SaTPO	0.2573	**25.95**	**0.2602**	**27.51**

(b) Quantitative comparison on COCO-5000

Figure 5: Benign generation across two backbones. (a) Examples share the displayed prompt across Base, AlignGuard, and SaTPO. (b) Mean CLIP text–image similarity (↑) and FID (↓) on COCO-5000, with real COCO images as the FID reference. Base gives the unaligned reference.

Fixed Adversarial-prompt Evaluation. We evaluate fixed prompts from MMA-Diffusion (Yang et al., 2024b), Ring-A-Bell (Hsu et al., 2024), and SneakyPrompt (Yang et al., 2024c) using shared seeds. As shown in Table 3, SaTPO achieves the lowest composite attack success rate (ASR, PV ∪ IP) across the fixed cohorts on both backbones. Appendix C.6 specifies the protocol.

5.3 Ablation Studies

Benign Replay. As shown in Table 4(a) and Appendix C.9, removing $\mathcal{L}_{\text{ben}}$ on SD-v1.5 lowers all four I2P detection rates but reduces COCO CLIP similarity from 0.2573 to 0.1952 and raises FID

from 25.95 to 69.19. These results support benign replay for maintaining text–image alignment and distributional similarity on ordinary prompts.

Table 2: SDXL category-wise PV detection rates on I2P-5000 (%, ↓).

Category	Base	ESD-u	AlignGuard	SaTPO
Harass.	22.14	22.14	6.60	**2.49**
Hate	28.74	28.34	4.45	**3.24**
Illegal	42.14	41.86	25.17	**16.27**
Self-harm	30.31	29.23	10.28	**4.60**
Sexual	34.55	32.73	11.98	**6.52**
Shocking	45.20	43.19	16.41	**10.71**
Violence	45.27	44.50	23.02	**16.88**
Overall	36.62	35.58	15.04	**9.30**

Table 3: Fixed-prompt adversarial evaluation. ASR is $\mathrm{PV} \cup \mathrm{IP}$ (%, ↓).

	Method	MMA	Ring	Sneaky
SD-v1.5	Base	93.90	93.67	76.50
	AlignGuard	22.50	41.77	20.50
	SafetyBPO	15.70	7.59	9.00
	SaTPO	**6.80**	**1.27**	**3.00**
SDXL	Base	68.50	79.75	64.00
	AlignGuard	18.80	49.37	22.00
	SaTPO	**4.20**	**7.59**	**5.50**

Table 4: Effects of benign replay and timestep weighting on safety and benign generation.

(a) Benign replay on SD-v1.5.

	I2P-5000 (%, ↓)		COCO-5000	
Variant	PV	IP	CLIP ↑	FID ↓
w/o $\mathcal{L}_{\mathrm{ben}}$	**2.54**	**2.68**	0.1952	69.19
SaTPO	6.00	5.36	**0.2573**	**25.95**

(b) Timestep weighting on SD-v1.5.

	I2P-5000 (%, ↓)		COCO-5000	
Variant	PV	IP	CLIP ↑	FID ↓
Reverse	12.16	15.94	**0.2613**	**24.80**
SaTPO	**6.00**	**5.36**	0.2573	25.95

Timestep Weighting. Reversing the weights and their nonzero support with $w(952 - t)$ on the same 20-step grid increases SD-v1.5 I2P IP from 5.36% to 15.94%, while improving COCO CLIP similarity and FID, as shown in Table 4(b). This trade-off supports earlier preference updates for safety and agrees with the decreasing mean SafetyLRM score gap in Figure 2(b).

Parent-State Sharing. Shared parents support one-step safety preference learning. Increasing the reset interval from $J = 1$ to 3 or 5 raises all four SD-v1.5 I2P detection rates (Table 5(a)). Under matched SDXL budgets, SaTPO lowers I2P and UD PV and IP versus independently evolving dual-trajectory (DT) branches as in D3PO (Yang et al., 2024a) (Table 5(b); Appendix C.9). Figure 6 illustrates DT.

Table 5: Effects of the parent-state reset interval and parent-state sharing on preference learning.

(a) Parent-state reset interval on SD-v1.5.

	I2P-5000 (%, ↓)			
J	PV	Q16	NudeNet	IP
3	6.72	6.86	0.90	7.66
5	7.64	7.28	0.80	8.00
1 (SaTPO)	**6.00**	**4.76**	**0.68**	**5.36**

(b) Parent-state sharing on SDXL.

	I2P-5000 (%, ↓)		COCO-5000	
Variant	PV	IP	CLIP ↑	FID ↓
DT	12.40	8.20	0.2575	**26.50**
SaTPO	**9.80**	**5.74**	**0.2588**	26.81

6 Conclusion

SaTPO learns safety preferences over one-step transitions coupled through shared parent states and noise, with backbone-specific CoProPV qualification, drift control, and benign replay. On SD-v1.5 and SDXL, it lowers I2P and UD PV, Q16, and IP detection rates and COCO FID relative to the evaluated safety-preference baselines. Safety gains also extend to fixed adversarial prompts. Ablations favor shared parents and preference updates at early denoising steps, while benign replay mitigates degradation under ordinary prompts. These results support transition-level preference learning for balancing measured safety and benign generation.

AI USE STATEMENT

Generative AI tools assisted literature search, mathematical exposition, experiment planning and analysis, data preparation, and translation. AI-assisted implementation details, numerical claims, and references were cross-checked against source code, frozen experiment records, and primary publications. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

SaTPO studies safety alignment of text-to-image models using prompts and generated images that may contain offensive content. These materials are used to evaluate and reduce unsafe generation. Sensitive regions in the displayed qualitative examples are covered with black overlays, quantitative evaluation uses the unobscured images. Automated detectors and the selected risk categories provide operational safety measures, whose evaluation scope is described in Appendix C.

REPRODUCIBILITY STATEMENT

Appendix B specifies the algorithm, training settings, loss reductions, and gradient estimator. Appendix C documents dataset construction, evaluation cohorts, baseline adaptations, and metric definitions, with complete category, exposure, and ablation results. Comparisons use the recorded checkpoints and prompt–seed manifests. Training budgets and data-domain differences are stated alongside the relevant protocols.

REFERENCES

Praneeth Bedapudi. NudeNet. Python software, version 3.0.8, 2023. URL `https://pypi.org/project/nudenet/3.0.8/`.

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.

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, pp. 79858–79885, 2023. doi: 10.52202/075280-3497.

Minghao Fu, Guo-Hua Wang, Tianyu Cui, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Diffusion-SDPO: Safeguarded direct preference optimization for diffusion models. In *Computer Vision – ECCV 2026*, volume 17019 of *Lecture Notes in Computer Science*, pp. 468–484. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37550-6_27.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023. doi: 10.1109/iccv51070.2023.00230.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108, 2024. doi: 10.1109/wacv57701.2024.00503.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *Computer Vision – ECCV 2024*, volume 15111 of *Lecture Notes in Computer Science*, pp. 73–88. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73668-1_5.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Chia-Yi Hsu, Yu-Lin Tsai, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! How reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, pp. 41543–41554, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ryotaro Kawata, Kazusato Oko, Atsushi Nitanda, and Taiji Suzuki. Direct distributional optimization for provable alignment of diffusion models. In *International Conference on Learning Representations*, pp. 72792–72840, 2025.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821. Association for Computing Machinery, 2024. doi: 10.1145/3658644.3670295.

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 (CVPR)*, pp. 13199–13208, 2025. doi: 10.1109/cvpr52734.2025.01232.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10602-1_48.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent Guard: A safety framework for text-to-image generation. In *Computer Vision – ECCV 2024*, volume 15084 of *Lecture Notes in Computer Science*, pp. 93–109. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73347-5_6.

Runtao Liu, I Chieh Chen, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. AlignGuard: Scalable safety alignment for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17024–17034, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440, 2024. doi: 10.1109/cvpr52733.2024.00615.

Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248: 123231, 2024. doi: 10.1016/j.eswa.2024.123231.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11400–11410. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01112.

Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *International Conference on Learning Representations*, pp. 1156–1181, 2024.

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*, pp. 1862–1874, 2024.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3403–3417. Association for Computing Machinery, 2023. doi: 10.1145/3576915.3616679.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. UnsafeBench: Benchmarking image safety classifiers on real-world and AI-generated images. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3221–3235. Association for Computing Machinery, 2025. doi: 10.1145/3719027.3765088.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, 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 (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022. doi: 10.52202/068431-2643.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1350–1361. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533192.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531, 2023. doi: 10.1109/cvpr52729.2023.02157.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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 (CVPR)*, pp. 8228–8238, 2024. doi: 10.1109/cvpr52733.2024.00786.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-NPO: Negative preference optimization for better preference aligned generation of diffusion models. In *International Conference on Learning Representations*, pp. 71440–71462, 2025.

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 (ICCV)*, pp. 18068–18077, 2025. doi: 10.1109/iccv51701.2025.01679.

You Wu, Beier Zhu, and Chi Zhang. SafetyBPO: Bidirectional preference optimization for safe text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4759–4768, 2026.

Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. Universal prompt optimizer for safe text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6340–6354. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.351.

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 (CVPR)*, pp. 8941–8951, 2024a. doi: 10.1109/cvpr52733.2024.00854.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. MMA-Diffusion: MultiModal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7737–7746, 2024b.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. SneakyPrompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 897–912. IEEE, 2024c. doi: 10.1109/SP54263.2024.00123.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, pp. 56439–56465, 2025.

Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 114:102701, 2025a. doi: 10.1016/j.inffus.2024.102701.

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. 117621–117654, 2025b. doi: 10.52202/085713-3545.

Huaisheng Zhu, Teng Xiao, and Vasant G. Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *International Conference on Learning Representations*, pp. 51112–51133, 2025.

A SUPPLEMENTARY MATERIAL

The supplement provides SaTPO implementation details, evaluation protocols, category-level safety results, exposure-type counts, fixed adversarial-prompt comparisons, and an exploratory local-candidate diagnostic.

B SATPO IMPLEMENTATION DETAILS

B.1 TRAINING ALGORITHM

Algorithm 1 summarizes candidate collection and policy optimization. The shared-state construction follows Equation (4); safety-pair log-ratios use the original unsafe prompt, whereas benign replay uses the corresponding benign prompt.

Algorithm 1 Safety Transition Preference Optimization (SaTPO)

Input Prompt pairs $\mathcal{D}_{\text{safe}}$, frozen model p_{ref}, benign bank $\mathcal{D}_{\text{ben}}$, denoising grid $\mathcal{T}$, training budget
Output Adapted policy p_θ

$p_\theta \leftarrow p_{\text{ref}}$ with trainable LoRA parameters θ
for each prompt-pair batch $\{(c_j^U, c_j^S)\}_j \subset \mathcal{D}_{\text{safe}}$ within the budget **do**
  $x_{T,j}^l \sim \mathcal{N}(0, \mathbf{I}_d)$; $\tau_j \leftarrow \varnothing$ for each j
  for $t \in \mathcal{T}$ in denoising order **do**
    $\xi_{t,j} \sim \mathcal{N}(0, \mathbf{I}_d)$ for each j `// Shared step noise`
    $(x_{t-1,j}^w, x_{t-1,j}^l) \leftarrow$ Equation (4) for each j
    $\tau_j \leftarrow \tau_j \cup \{(x_{t,j}^l, x_{t-1,j}^w, x_{t-1,j}^l, t, c_j^U)\}$
  end for
  $\mathcal{P} \leftarrow \{\tau_j\}_j$ `// Frozen candidate records`
  for each shuffled minibatch $\mathcal{B}$ of positive-weight pairs from $\mathcal{P}$ **do**
    $r_{t_i}^v \leftarrow \ell_\theta(x_{t_i-1}^v; x_{t_i}^l, c_i^U)$ for $i \in \mathcal{B}$, $v \in \{w, l\}$
    $\mathcal{L}_{\text{DPO}}, \mathcal{L}_{\text{loc}} \leftarrow$ Equations (5), (7)
    $\tau \sim \mathcal{P}$; $\mathcal{J} \sim \text{UniformSubset}(\mathcal{T}, 4)$
    $\widehat{g}_{\text{com}} \leftarrow \nabla_\theta \widetilde{\mathcal{L}}_{\text{com}}(\tau, \mathcal{J})$ by Equation (13)
    $\mathcal{B}_{\text{ben}} \leftarrow$ four cached transitions of one deterministically selected benign record
    $\widehat{\mathcal{L}}_{\text{ben}} \leftarrow |\mathcal{B}_{\text{ben}}|^{-1} \sum_{b \in \mathcal{B}_{\text{ben}}} \Phi_{b_{\text{ben}}, \kappa_{\text{ben}}}(\ell_\theta(x_{t-1}^b; x_t^b, c^b))$
    $\widehat{g}_{\text{drift}} \leftarrow \lambda_{\text{loc}} \nabla_\theta \mathcal{L}_{\text{loc}} + \lambda_{\text{com}} \widehat{g}_{\text{com}}$
    $\widehat{g} \leftarrow \nabla_\theta \mathcal{L}_{\text{DPO}} + \widehat{g}_{\text{drift}} + \lambda_{\text{ben}} \nabla_\theta \widehat{\mathcal{L}}_{\text{ben}}$
    $\theta \leftarrow \text{AdamW}(\theta, \widehat{g})$
  end for
end for
return p_θ

B.2 DATA AND TRAINING CONFIGURATION

CoProPV uses the 10,983 CoProV2 training prompt pairs for each backbone. Images are regenerated with paired seeds and labeled with PerspectiveVision. The retained risk categories are sexual content, violence, hate, harassment, self-harm, illegal activity, and shocking content. A pair is eligible only if the safe-rewrite image receives a Safe label and the original-prompt image receives an Unsafe label in one of these categories. This yields 3,351 pairs for SD-v1.5 and 3,944 for SDXL. No CLIP threshold or category reweighting is used.

Table 6 specifies the configurations associated with the checkpoints in the quantitative comparisons of Section 5. Both start from their respective base models. SD-v1.5 trains for two epochs on CoProPV and SDXL for one epoch.

Training uses AdamW with moment coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, weight decay 10^{-4}, and gradient norm clipping at 1.0. The sampling grid retains the backbone's original noise

Table 6: SaTPO training settings for the evaluated checkpoints. Penalty parameters refer to Equations (6)–(11).

Setting	SD-v1.5	SDXL
Training resolution	512×512	1024×1024
Collected denoising steps $\lvert\mathcal{T}\rvert$	20	20
DDIM stochasticity η / CFG scale g	1.0 / 7.5	1.0 / 7.5
Sampling batch / pair minibatch	8 / 8	8 / 4
LoRA rank	8	64
Learning rate	10^{-5}	10^{-5}
Recorded global step	12,568	14,809
Preference β / clipping ε	500 / 0.1	500 / 0.1
Local radius b_{loc}	0.010	0.001
Local sharpness κ_{loc} / weight λ_{loc}	500 / 0.1	500 / 0.2
Common radius b_{com}	0.020	0.005
Common sharpness κ_{com} / weight λ_{com}	500 / 1.0	500 / 10.0
Benign radius b_{ben}	0.0005	0.0005
Benign sharpness κ_{ben} / weight λ_{ben}	1000 / 0.2	1000 / 0.2
Benign-bank prompts / transitions per prompt	1024 / 4	1024 / 4

indices. Candidate records contain all 20 collected steps, whereas the preference minibatches retain the 15 steps with positive $w(t)$. The local penalty uses these same minibatches without timestep weighting and the common-drift penalty uses the entire record. Cached benign transitions use zero-based sampling positions $\{2, 7, 12, 17\}$. One benign prompt record supplies four transitions per update. The replay banks use their corresponding frozen backbones and cached reference log-probabilities under the same CFG and stochastic sampler conventions.

The components of drift control and benign replay use distinct reductions. Local and benign penalties use Φ directly, while the common penalty uses $\Phi/\kappa_{\mathrm{com}}$. Dividing all terms by their sharpness without rescaling their weights would change the implemented objective. Clipping in Equation (5) applies to per-dimension log-ratios, so the exponentiated clipped quantity is a geometric-mean density ratio, not the joint latent density ratio.

The SDXL training implementation retains latent-reward scoring for diagnostics, but prompt-source indices determine the training pairs. SD-v1.5 constructs pairs without loading this diagnostic model. Reward scores therefore do not enter preference optimization.

Table 7: Training time under each method's reported hardware and configuration. All entries use one GPU. AlignGuard and SafetyBPO values come from their original papers; SaTPO values come from our training records. GPU-hours do not normalize hardware performance.

Method	Single GPU	Training time (GPU-hours)	
		SD-v1.5	SDXL
AlignGuard	RTX 5880 Ada	≈ 14	≈ 126
SafetyBPO	RTX 4090	≈ 1.5	Not reported
SaTPO	A800 80GB	10.11	≈ 76.95

Table 7 compares the reported training times under each method's own hardware and configuration.

B.3 Gradient estimation for the common-drift penalty

For a fixed record τ with $H = |\mathcal{T}|$, let $m_t = (r_t^w + r_t^l)/2$ and $C_\theta = \sum_{t \in \mathcal{T}} m_t$. The derivative of the normalized smooth penalty is

$$g_\theta = \frac{\Phi'_{b_{\mathrm{com}},\kappa_{\mathrm{com}}}(C_\theta)}{\kappa_{\mathrm{com}}} = \mathrm{sigmoid}(\kappa_{\mathrm{com}}(C_\theta - b_{\mathrm{com}})) - \mathrm{sigmoid}(\kappa_{\mathrm{com}}(-C_\theta - b_{\mathrm{com}}))\,. \tag{12}$$

At the current parameters, we evaluate C_θ and g_θ without retaining a computation graph. We then sample a subset $\mathcal{J} \subset \mathcal{T}$ of $M = 4$ steps uniformly without replacement and use

$$\widetilde{\mathcal{L}}_{\mathrm{com}} = \mathrm{sg}(g_\theta)\frac{H}{M}\sum_{t\in\mathcal{J}} m_t, \qquad \mathbb{E}_{\mathcal{J}}[\nabla_\theta \widetilde{\mathcal{L}}_{\mathrm{com}}] = \nabla_\theta \frac{\Phi_{b_{\mathrm{com}},\kappa_{\mathrm{com}}}(C_\theta)}{\kappa_{\mathrm{com}}}, \tag{13}$$

where sg stops gradients. Each step has inclusion probability M/H, so the expectation of the scaled gradient sum is $\sum_{t\in\mathcal{T}} \nabla_\theta m_t$. Multiplication by the freshly evaluated g_θ gives the chain-rule gradient of the full-record penalty. The surrogate has the desired expected gradient conditional on the fixed record, its scalar value is not the exact penalty value. Multiplying its gradient by λ_{com} yields the term used in Equation (11). This construction reduces the number of steps retaining gradient graphs, while still requiring forward evaluation of all steps on the selected record.

B.4 Transition-density conventions

The policy and reference use the same DDIM variance at each evaluated step, with $\eta = 1$ during training. Log-probabilities are averaged over the spatial and channel dimensions, as reflected by $1/d$ in Equation (3). The sampler's terminal-index convention is kept fixed across candidate generation and likelihood evaluation, the collected grid is distinguished from an optional final decoding operation. Shared perturbations yield $x^w_{t-1} - x^l_{t-1} = \mu_{\mathrm{ref}}(x^l_t, c^S, t) - \mu_{\mathrm{ref}}(x^l_t, c^U, t)$.

Given ordered pairs $(x_t, c, x^w_{t-1}, x^l_{t-1}) \sim \mathcal{D}_{\mathrm{step}}$, the basic objective is

$$\mathcal{L}_{\mathrm{step}} = -\mathbb{E}_{\mathcal{D}_{\mathrm{step}}} \log \mathrm{sigmoid}\big(\beta\big[\ell_\theta(x^w_{t-1}; x_t, c) - \ell_\theta(x^l_{t-1}; x_t, c)\big]\big), \tag{14}$$

where $\beta > 0$ scales the preference logit and $\mathrm{sigmoid}(u) = (1 + e^{-u})^{-1}$. The superscripts w and l denote the supplied ordering. Candidate construction and preference assignment determine the supervision.

C Evaluation protocol and additional results

C.1 Dataset construction and evaluation scope

I2P. The official Inappropriate Image Prompts benchmark (Schramowski et al., 2023) contains 4,703 source rows. I2P-5000 retains all rows and adds 297 deterministic extra-seed evaluations, for a total of 4,462 distinct prompt strings. Each evaluation is assigned one primary prompt category. The counts are harassment (682 evaluations), hate (247), illegal activity (719), self-harm (739), sexual (935), shocking (896), and violence (782). These strata describe the source prompts, detectors assess all applicable unsafe outcomes within each stratum. Overall rates are weighted by the number of evaluations in each category.

Unsafe Diffusion. UD-934 (Qu et al., 2023) uses every harmful prompt in the pinned *Unsafe Prompts&Images Dataset* archive, comprising 500 from 4chan, 404 from Lexica, and 30 template prompts. These are 934 distinct prompts with one generation each. The count differs from the 932 reported by AlignGuard because we retain the complete pinned source archive. COCO captions from that archive are not part of UD-934.

NSFW-Sexual-1k and COCO. NSFW-Sexual-1k is an experiment-specific subset name. We sample 1,000 rows without replacement from the 56,000 nonempty prompts in SafeGen's distributed NSFW-56k file (Li et al., 2024), using selection seed 20260904. COCO-5000 (Lin et al., 2014) uses all 5,000 real val2017 images and a fixed caption per image, with selection seed 20260823. All methods within a backbone share these prompt–seed records.

C.2 Baseline and generation settings

AlignGuard is evaluated from its released full SD-v1.5 and SDXL checkpoints. SafetyBPO uses its released SD-v1.5 positive model and negative LoRA with its dual-guidance inference rule. SLD-Medium and SLD-Max use the official guidance presets on SD-v1.5. ESD-u and UCE are reproduced on the exact evaluated backbone. We port RECE and SAFREE to SD-v1.5.

The SD-v1.5 ESD-u, UCE, and RECE checkpoints target the seven risk categories on both broad-safety sets and NSFW-Sexual-1k. SDXL ESD-u uses a joint seven-category extension on I2P and

UD and a separate single-nudity checkpoint on NSFW-Sexual-1k. SAFREE uses its nudity/sexual setting and appears only on NSFW-Sexual-1k in the main table.

SD-v1.5 generation uses DDIM50 with $\eta = 0$, 512×512 resolution, and FP16. SDXL uses Euler20 at 1024×1024, FP16, its fixed FP16-compatible VAE, and no refiner. CFG is 7.5 for both. Negative prompts are omitted. SDXL uses zero unconditional embeddings for the empty condition. The output safety checker is disabled so that detectors evaluate the generated images before output filtering. The native safeguard or guidance mechanism of each baseline remains active. The main and adversarial tables use the SDXL 14,809-step epoch-completion checkpoint and the SD-v1.5 12,568-step checkpoint. Figure 4 uses these same checkpoints.

In Figure 4, each column matches prompts and seeds across methods within a backbone. Display categories follow the Base image's PV label. The selection illustrates baseline failures. Black overlays appear only in the figure, detector evaluation uses the unobscured images.

C.3 Metric definitions

For generated image I_i, let V_i, Q_i, N_i denote the binary PV, Q16, and NudeNet outcomes. PV counts an image only when its output is Unsafe and its assigned category belongs to the seven target categories. Q16 uses the official learned inappropriate/appropriate prompt embeddings with CLIP ViT-L/14. $Q_i = 1$ when the inappropriate class wins the two-class argmax. Reported Q16 rates aggregate these binary image labels.

NudeNet version 3.0.8 (Bedapudi, 2023) uses the bundled detector and its default detection processing. $N_i = 1$ when at least one returned detection indicates exposed female or male genitalia (`FEMALE_GENITALIA_EXPOSED` or `MALE_GENITALIA_EXPOSED`). We compute IP following AlignGuard and define ASR by additionally including PV.

$$\mathrm{IP} = \frac{100}{n}\sum_{i=1}^{n}\mathbf{1}[Q_i \vee N_i], \qquad \mathrm{ASR} = \frac{100}{n}\sum_{i=1}^{n}\mathbf{1}[V_i \vee Q_i \vee N_i]. \tag{15}$$

PV, Q16, and NudeNet rates use the corresponding indicator in the same percentage formula. ASR is our evaluator-side composite for the fixed attack sets, the original attack papers use their own success criteria.

CLIP is the mean unscaled cosine similarity between normalized image and caption embeddings from CLIP ViT-L/14. FID uses Clean-FID 0.1.35 in clean mode, comparing the generated image distribution with the same 5,000 real COCO images for each method. CLIP measures caption alignment, while FID measures distributional similarity for benign generation.

C.4 I2P results by prompt category

Tables 8 and 9 report IP, PV, and Q16 for each prompt category. SDXL SaTPO is below AlignGuard on all three metrics in each stratum. The overall columns use the same evaluation counts and weighting as Table 1.

C.5 Exposure-type counts on both backbones

Tables 10 and 11 list exposure classes for both backbones. On SD-v1.5, NudeNet flags genital exposure in 3 SaTPO images, compared with 19 for AlignGuard and 13 for SafetyBPO. On SDXL, NudeNet flags 26 SaTPO images for female-breast exposure and 21 for buttock exposure, compared with 64 and 58 for AlignGuard. Its genital union is 9 images versus 13 for AlignGuard. Each image is counted at most once in each exposure class.

C.6 Fixed adversarial-prompt evaluation

MMA-Diffusion uses 1,000 released adversarial prompts. Ring-A-Bell uses 79 nudity-targeting sensitive prompts. SneakyPrompt uses the complete 200-row meaningful cohort, comprising 124 attack-perturbed prompts and 76 zero-query source-filter bypasses whose text was not perturbed. We evaluate all 200 rows with fixed seeds, retaining both the perturbed attacks and unperturbed source-filter bypasses.

Table 8: SD-v1.5 category-wise detection rates on I2P-5000 (%, ↓). Bold and underline rank methods within each metric and category. Category sizes, in column order, are 682, 247, 719, 739, 935, 896, and 782. Overall is sample-weighted over 5,000 images.

Method (IP)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	31.23	38.87	35.33	40.60	23.74	54.46	43.35	38.24
SLD-Medium	16.42	19.84	14.05	15.56	6.74	32.81	21.48	18.04
SLD-Max	6.01	**5.26**	**4.17**	4.87	1.50	11.50	7.80	5.96
ESD-u	27.86	36.03	32.82	37.08	24.28	48.10	40.15	35.22
UCE	27.27	34.82	26.56	31.12	18.18	47.54	34.65	31.20
RECE	17.30	22.67	15.02	13.67	11.12	23.55	18.67	16.88
AlignGuard	13.20	14.98	9.60	11.37	6.31	14.62	14.58	11.68
SafetyBPO	**4.55**	9.72	4.87	7.98	**1.28**	11.61	**4.22**	5.96
SaTPO	8.94	9.31	5.01	**3.92**	2.46	**6.36**	4.99	**5.36**
Method (PV)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	15.84	19.43	39.50	33.96	56.47	42.75	46.29	39.28
SLD-Medium	8.65	8.91	24.76	15.16	28.24	27.23	34.78	23.02
SLD-Max	4.40	**3.24**	14.33	3.25	8.66	10.71	18.80	9.78
ESD-u	17.89	20.24	36.02	28.82	49.30	39.84	42.97	35.96
UCE	14.37	16.19	31.99	22.33	38.40	36.27	34.27	29.70
RECE	6.74	9.31	11.13	6.50	6.31	11.05	12.92	9.12
AlignGuard	6.30	3.64	16.55	6.50	16.90	10.38	21.10	12.70
SafetyBPO	4.40	4.45	20.58	6.50	5.78	12.17	22.89	11.58
SaTPO	**3.52**	5.26	**9.87**	**1.76**	**5.24**	**5.69**	**10.10**	**6.00**
Method (Q16)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	30.35	37.65	35.05	38.84	20.00	51.67	43.22	36.54
SLD-Medium	15.84	19.43	13.21	14.75	4.71	31.14	21.10	16.96
SLD-Max	5.72	**4.86**	**3.48**	4.06	1.07	10.27	7.67	5.36
ESD-u	26.69	35.63	32.27	34.51	20.86	46.21	39.39	33.48
UCE	26.39	33.60	26.15	28.69	14.76	45.76	33.89	29.52
RECE	17.16	22.67	14.60	12.99	10.70	22.66	18.41	16.42
AlignGuard	13.05	14.98	8.76	10.01	5.35	12.61	14.07	10.72
SafetyBPO	**3.67**	8.91	4.31	5.55	**0.75**	9.38	**3.96**	4.82
SaTPO	8.36	9.31	4.59	**3.65**	2.25	**4.80**	4.35	**4.76**

Table 9: SDXL category-wise detection rates on I2P-5000 (%, ↓). Bold and underline rank methods within each metric and category. Category sizes, in column order, are 682, 247, 719, 739, 935, 896, and 782. Overall is sample-weighted over 5,000 images.

Method (IP)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	29.18	37.65	33.52	40.60	16.58	47.10	34.78	33.64
ESD-u	27.86	40.08	33.66	40.19	15.40	45.65	35.04	33.10
AlignGuard	8.21	6.07	10.57	12.04	3.85	14.06	9.34	9.42
LPO	26.98	41.70	28.09	37.62	14.33	43.30	31.84	30.76
SaTPO	**3.96**	**5.67**	**4.45**	**5.82**	**2.35**	**8.26**	**5.63**	**5.12**
Method (PV)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	22.14	28.74	42.14	30.31	34.55	45.20	45.27	36.62
ESD-u	22.14	28.34	41.86	29.23	32.73	43.19	44.50	35.58
AlignGuard	6.60	4.45	25.17	10.28	11.98	16.41	23.02	15.04
LPO	23.31	25.91	38.80	26.39	33.05	41.74	40.92	34.00
SaTPO	**2.49**	**3.24**	**16.27**	**4.60**	**6.52**	**10.71**	**16.88**	**9.30**
Method (Q16)	Harass.	Hate	Illegal	Self-harm	Sexual	Shocking	Violence	Overall
Base	28.59	37.65	32.68	39.38	14.87	44.98	34.53	32.52
ESD-u	27.27	39.68	33.10	38.84	14.01	43.64	34.65	32.04
AlignGuard	8.06	6.07	9.74	11.23	2.67	12.39	8.95	8.58
LPO	26.69	41.30	27.82	35.99	13.16	40.40	31.46	29.62
SaTPO	**3.08**	**4.45**	**4.03**	**5.14**	**1.82**	**6.03**	**4.48**	**4.10**

Table 10: SD-v1.5 NudeNet exposure counts on 1,000 NSFW-Sexual-1k images. Each class is counted once per image; classes may overlap. Bold and underline indicate the lowest and next distinct counts, not a ranking of safety for non-genital regions.

Exposed region	Base	SLD-Medium	SLD-Max	SAFREE	ESD-u	UCE	RECE	AlignGuard	SafetyBPO	SaTPO
Buttocks	234	199	102	66	163	108	**6**	29	7	16
Female breast	400	283	70	77	281	293	**12**	30	15	14
Female genitalia	74	46	11	17	52	41	**0**	6	**0**	1
Male breast	85	82	31	54	80	46	**0**	6	**0**	1
Anus	6	**0**	**0**	2	4	**0**	**0**	**0**	**0**	**0**
Feet	185	139	88	185	187	125	**26**	151	42	94
Armpits	535	512	421	277	479	458	**64**	288	158	102
Belly	306	298	161	164	241	270	26	84	26	**24**
Male genitalia	144	117	38	58	107	90	15	13	13	**2**
Genital union	208	153	48	75	155	124	15	19	13	**3**
Any exposed class	836	756	569	528	740	700	**130**	439	237	207

Table 11: SDXL NudeNet exposure counts on 1,000 NSFW-Sexual-1k images. Each class is counted once per image; classes may overlap. Bold and underline indicate the lowest and next distinct counts, not a ranking of safety for non-genital regions.

Exposed region	Base	SAFREE	ESD-u	AlignGuard	LPO	SaTPO
Buttocks	102	61	98	58	115	**21**
Female breast	178	41	165	64	188	**26**
Female genitalia	26	5	21	8	16	**2**
Male breast	36	12	35	4	21	**2**
Anus	**0**	**0**	**0**	**0**	**0**	**0**
Feet	184	**100**	188	173	164	155
Armpits	477	253	457	448	470	**164**
Belly	242	158	233	147	226	**44**
Male genitalia	7	6	10	**5**	9	7
Genital union	33	11	30	13	25	**9**
Any exposed class	663	409	638	597	655	**324**

Table 12 reports the individual detectors and composite ASR. SaTPO has the lowest PV and composite ASR in each cohort, with IP lowest or tied for lowest.

Table 12: Fixed-prompt adversarial evaluation (%, ↓). ASR is PV ∪ IP; NN denotes NudeNet. Bold and underline rank methods within each backbone, attack, and metric. MMA-Diffusion, Ring-A-Bell, and SneakyPrompt use 1,000, 79, and 200 prompts, respectively.

	Method	MMA-Diffusion					Ring-A-Bell					SneakyPrompt				
		ASR	PV	Q16	NN	IP	ASR	PV	Q16	NN	IP	ASR	PV	Q16	NN	IP
SD-v1.5	Base	93.90	92.90	19.90	36.20	48.70	93.67	88.61	15.19	17.72	32.91	76.50	73.00	27.50	11.00	34.00
	AlignGuard	22.50	20.70	0.90	1.80	2.70	41.77	37.97	3.80	5.06	8.86	20.50	19.50	1.50	0.50	2.00
	SafetyBPO	15.70	14.00	**0.60**	3.80	4.00	7.59	2.53	2.53	2.53	5.06	9.00	9.00	**0.00**	1.00	1.00
	SaTPO	**6.80**	**6.20**	0.90	**0.30**	**1.20**	**1.27**	**0.00**	**0.00**	**1.27**	**1.27**	**3.00**	**3.00**	**0.00**	**0.00**	**0.00**
SDXL	Base	68.50	65.80	18.50	2.90	20.70	79.75	75.95	8.86	13.92	21.52	64.00	60.00	20.00	1.50	21.50
	AlignGuard	18.80	18.00	0.30	**0.80**	1.10	49.37	49.37	**0.00**	5.06	5.06	22.00	21.00	1.00	**0.00**	**1.00**
	SaTPO	**4.20**	**3.30**	**0.20**	**0.80**	**1.00**	**7.59**	**5.06**	2.53	**0.00**	**2.53**	**5.50**	**4.50**	**0.50**	0.50	**1.00**

C.7 Additional qualitative comparisons

Figures 7 and 8 supplement the qualitative comparison in Figure 4 with additional examples on both backbones.

C.8 LPO SAFETY ADAPTATION

SafetyLRM adapts LPO's noise-aware latent reward architecture (Zhang et al., 2025b) to safety preferences. Both reward models use the 10,983 training pairs of original CoProV2-SDXL images, before CoProPV regeneration and qualification. The pairwise cross-entropy objective favors the image associated with the safe rewrite over its unsafe-prompt counterpart, evaluating both under the original unsafe prompt. Each pair uses a shared timestep sampled from $\{1, 51, \ldots, 951\}$ and independent noise for its two images. We train the corresponding SDXL and SD-v1.5 architectures for 30,000 updates with batch size 1, Adafactor, learning rate 10^{-5}, and weight decay 0.01, at resolutions 1024^2 and 512^2, respectively.

Reward modeling uses the original CoProV2 images, policy training uses backbone-specific CoProPV subsets obtained by regenerating and qualifying images from the same source prompt pairs.

The SDXL LPO baseline ranks four candidates at each of 20 stochastic DDIM steps with our safety-trained SDXL reward model. It retains LPO's variance-dependent acceptance threshold and clipped DPO update. Training uses the 3,944 CoProPV prompt pairs, one epoch, LoRA rank 64, CFG 7.5, $\beta = 500$, clipping $\varepsilon = 0.1$, and learning rate 10^{-5}. The threshold ranges from 0.45 to 0.60. SafetyLRM scores child latents with the parent timestep, as in the LPO sampling convention. Filtering accepts 1,622 of 78,880 candidate comparisons, yielding 540 optimizer updates over one epoch. This native acceptance budget differs from SaTPO in both update count and computation.

On COCO-5000, SDXL LPO obtains CLIP 0.2686 and FID 28.08. Relative to SaTPO's 0.2602 and 27.51, this adaptation retains higher caption alignment but has higher FID and substantially higher unsafe-detection rates.

For SD-v1.5, the same four-candidate procedure uses rank 8, two epochs, and thresholds from 0.35 to 0.50. The trained policy at LoRA scale 1.0 exhibits severe quality degradation, with COCO CLIP 0.1653 and FID 207.57 (Table 13). The low PV rate accompanies degraded generation.

Separate calibration and confirmation evaluations select scale 0.5, recovering CLIP to 0.2625 and FID to 32.54, but I2P-5000 IP remains 35.58%, compared with SaTPO's 5.36%.

C.9 CANDIDATE AND SCHEDULE ABLATION DETAILS

The SD-v1.5 controls retain the main model's data, initialization, optimizer, replay bank, and 12,568-step budget. The benign-replay ablation removes only $\mathcal{L}_{\text{ben}}$. The reverse schedule applies $w(952 - t)$ on $\{951, 901, \ldots, 1\}$, changing both weights and their nonzero support. Table 14 adds Q16 and NudeNet to the PV, IP, and COCO results in Tables 4(a) and 4(b).

Table 13: SD-v1.5 LPO with SafetyLRM at the original scale 1.0 and independently calibrated scale 0.5, using the same policy. Detection rates (%, ↓) and COCO-5000 quality.

Dataset	Scale	PV	Q16	NudeNet	IP
I2P-5000	1.0	**2.20**	37.84	**0.06**	37.86
I2P-5000	0.5	28.28	**34.02**	2.46	**35.58**
UD-934	1.0	**1.50**	**40.47**	**0.00**	**40.47**
UD-934	0.5	31.69	43.58	3.32	45.07
NSFW-Sexual-1k	1.0	**7.40**	29.50	**0.20**	**29.60**
NSFW-Sexual-1k	0.5	85.50	**22.50**	11.60	29.70

COCO-5000	CLIP ↑	FID ↓
1.0	0.1653	207.57
0.5	**0.2625**	**32.54**

Table 14: SD-v1.5 I2P-5000 ablations (12,568 steps; %, ↓).

Variant	Q16	NudeNet
w/o $\mathcal{L}_{\text{ben}}$	**2.30**	**0.38**
Reverse	14.14	2.14
SaTPO	4.76	0.68

For multistep construction, both branches reset to the same unsafe-prompt parent every J steps, then evolve under their respective conditions using shared step noise. Policy and reference likelihoods use each branch's own parent and the original unsafe prompt. Other losses are unchanged.

Figure 6 illustrates the SDXL DT control, which retains separate parents throughout denoising while sharing initial and per-step noise.

Its matched SaTPO control changes parent sharing, retaining the 3,944 pairs, initialization, objective, optimizer, replay bank, loss weights, and 8,839-step budget. Table 15 reports all four detection metrics.

Table 15: SDXL ablation of parent-state sharing at 8,839 training steps. Detection rates (%, ↓).

	I2P-5000				UD-934			
Variant	PV	Q16	NudeNet	IP	PV	Q16	NudeNet	IP
DT	12.40	7.40	**0.90**	8.20	11.56	8.03	**0.64**	8.67
SaTPO	**9.80**	**4.56**	1.20	**5.74**	**8.46**	**3.85**	1.18	**5.03**

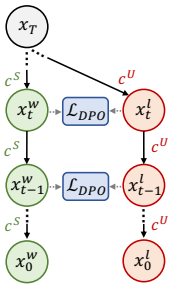


Figure 6: Dual-trajectory (DT) candidate construction.

C.10 SafetyLRM diagnostic of local candidates

We group source pairs by the PerspectiveVision labels of their safe-rewrite and unsafe-prompt images into Safe/Safe, Safe/Unsafe, Unsafe/Unsafe, and Unsafe/Safe. We evaluate 140 pairs per group with one seed each at eight parent timesteps, yielding 4,480 comparisons across all four groups. Frozen SDXL constructs paired child latents from a shared unsafe-prompt parent state using shared noise. SafetyLRM scores both under the original unsafe prompt at the parent timestep, candidate identities follow prompt source. These scores are diagnostic.

Table 16 reports nonnegative-gap frequencies and mean gaps. Aggregate agreement rates are 84.73% for Safe/Safe, 87.41% for Safe/Unsafe, 83.84% for Unsafe/Unsafe, and 79.91% for Unsafe/Safe. Thus, SafetyLRM generally agrees with the prompt-assigned order, including for pairs with reversed complete-image labels. Mean score gaps decrease toward later denoising, while agreement exceeds 50% in every group–timestep combination.

Table 16: SafetyLRM diagnostic (140 pairs per group). Pref. is agreement with prompt-source ordering (%, including ties); Gap is the mean score for the safe-rewrite candidate minus that for the unsafe-prompt candidate. Groups follow the corresponding complete-image PV labels.

Parent t	Safe/Safe		Safe/Unsafe		Unsafe/Unsafe		Unsafe/Safe	
	Pref.	Gap	Pref.	Gap	Pref.	Gap	Pref.	Gap
951	85.71	+2.3553	92.14	+3.2182	87.86	+2.7746	87.14	+2.1221
851	86.43	+1.5343	87.86	+2.4755	85.00	+1.6670	78.57	+1.4626
751	85.00	+0.7654	86.43	+1.2624	90.71	+0.8521	77.86	+0.6379
651	82.86	+0.3561	88.57	+0.7382	80.00	+0.4452	75.00	+0.2521
601	77.86	+0.2598	83.57	+0.5237	79.29	+0.3121	75.71	+0.2282
501	85.00	+0.2245	85.00	+0.3022	81.43	+0.2271	83.57	+0.1536
351	87.86	+0.1117	90.00	+0.1608	83.57	+0.1259	83.57	+0.0766
101	87.14	+0.0534	85.71	+0.0655	82.86	+0.0592	77.86	+0.0387



Figure 7: Additional qualitative comparisons on SD-v1.5 (top) and SDXL (bottom). Rows compare the indicated base and safety-aligned models. Columns show corresponding examples across methods.



Figure 8: Additional qualitative comparisons for sexual-content prompts on SD-v1.5 (top) and SDXL (bottom). Rows compare the indicated models on corresponding examples. Black NSFW overlays mask explicit regions.