DELTAPO: COUNTERFACTUALLY VERIFIED PREFERENCE OPTIMIZATION FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization for text-to-image generation typically relies on image-level labels that indicate which output is preferred, but not why. When candidate images differ in several respects, such supervision can mix better prompt satisfaction with incidental visual differences. We introduce DELTAPO, a preference optimization framework that tests the semantic basis of a preference through targeted image interventions. DELTAPO decomposes a prompt into verifiable requirements and identifies a semantic delta between matched candidates, together with the image regions that support it. It then tests this delta by repairing the corresponding violation in the rejected image and introducing the same violation into the preferred image. A delta is retained only when at least one intervention is valid and every valid intervention weakens the original preference while preserving non-target requirements and visual quality. The verified delta is then used to select training pairs and localize preference optimization to the implicated regions, while training remains conditioned on the original prompt and original image pair. Using SDXL as the base generator, DELTAPO improves compositional alignment on GenEval, T2I-CompBench, DPG-Bench, and GenEval 2, including gains from 0.53 to 0.64 on GenEval and from 73.38 to 79.97 on DPG-Bench. Ablations show that counterfactual verification provides gains beyond edit validity alone, while localized optimization further improves performance.

1 INTRODUCTION

Modern text-to-image models, including diffusion models (Esser et al., 2024; Chen et al., 2024) and unified multimodal models (Xie et al., 2024; 2025; Chen et al., 2025), can generate visually convincing scenes while still missing explicit constraints in the prompt. A prompt such as "three red apples to the left of a blue bowl" requires the correct objects, counts, attributes, and spatial relation to hold at the same time. An image may look realistic while violating any one of these requirements. Improving compositional generation therefore requires supervision that distinguishes prompt satisfaction from incidental visual quality.

Preference optimization learns from comparisons between preferred and rejected outputs. Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its adaptation to diffusion models (Wallace et al., 2023) provide objectives for such comparisons, while reward models such as ImageReward and PickScore (Xu et al., 2023; Kirstain et al., 2023) can help construct preference pairs at scale. However, an image-level preference says which output is better, not which visual difference is responsible for that preference. When two images differ in both prompt-relevant content and incidental appearance, optimizing the whole comparison can mix these signals.

Recent methods provide more fine-grained preference supervision. SILMM (Qu et al., 2024) derives feedback from decomposed visual questions; OSPO (Oh et al., 2026) constructs object-centric preference pairs and applies object-weighted optimization; BiDPO (Liu et al., 2026b) combines edited preference data with region-aware bimodal optimization; and RationalRewards (Wang et al., 2026) uses structured critiques for scoring and refinement. These methods make preference supervision more structured or localized. What remains missing is a way to test whether an identified semantic difference actually contributes to the observed preference.

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.

Consider two images of the apple-and-bowl prompt. One satisfies the left–right relation and is preferred, but also has more attractive lighting. A critic may attribute the preference to the relation, although the pair alone does not establish how much that relation contributes to the decision. If it does contribute, repairing the relation in the rejected image should reduce the preference gap, while introducing the same violation into the preferred image should weaken its advantage. Such a test is meaningful only if the remaining prompt requirements and visual quality stay unchanged. This motivates our central idea: *test a proposed semantic reason for a preference by intervening on the corresponding visual difference*.

We introduce DELTAPO, a framework that turns proposed semantic explanations of image preferences into counterfactually verified supervision. After decomposing each prompt into verifiable requirements, DELTAPO identifies candidate pairs matched on non-target requirements and proposes a *semantic delta*: a difference in requirement satisfaction grounded in visual evidence and localized image regions. We test this explanation by repairing the corresponding violation in the rejected image and introducing it into the preferred image. A delta is retained only if at least one intervention changes the intended requirement while preserving non-target requirements and visual quality, and every such valid intervention weakens the original preference by a fixed margin. The verified delta then determines both *which pairs to learn from* and *where to emphasize their supervision*, while a reference anchor discourages prediction drift outside the implicated regions. Crucially, counterfactual images serve as verification probes rather than replacement training data: optimization retains the original prompt and unedited image pair. The trained generator requires no critic or editor at inference time, while the same verification principle can optionally gate targeted repair.

Experiments with SDXL (Podell et al., 2023) on GenEval, T2I-CompBench, DPG-Bench, and GenEval 2 show consistent improvements in compositional alignment, spanning attribute binding, object composition, relations, and dense-prompt following. Beyond aggregate gains, our ablations distinguish *edit validity* from *preference relevance*: testing the preference response improves learning even after invalid interventions have been removed. The optimization analysis further shows that spatial emphasis alone is not sufficient for consistent gains: delta-grounded localization combined with reference anchoring achieves the strongest results on both ablation benchmarks and outperforms the random-mask control. In a separate inference-time study, verified repair continues to benefit from larger attempt budgets over the tested range, whereas ungated repair saturates, highlighting the importance of selective acceptance rather than additional editing alone.

2 RELATED WORK

Preference optimization for text-to-image generation. Direct Preference Optimization (DPO) (Rafailov et al., 2023) learns from preferred and rejected samples, and Diffusion-DPO (Wallace et al., 2023) adapts this objective to image generation. Reward-based fine-tuning has also been explored through DDPO (Black et al., 2024) and DPOK (Fan et al., 2023). ImageReward and PickScore (Xu et al., 2023; Kirstain et al., 2023) learn preference scores from human feedback, while PromptEcho (Liu et al., 2026a) constructs an annotation-free reward from the image-conditioned likelihood of the original prompt under a frozen vision–language model. These methods provide objectives or feedback signals for improving generated outputs. DELTAPO focuses on a complementary issue: whether the semantic difference used to justify an individual preference can be verified before that comparison contributes to training.

Fine-grained supervision and preference data. Question-based evaluation methods such as TIFA (Hu et al., 2023) and GenEval 2 (Kamath et al., 2025) assess whether individual prompt requirements are reflected in a generated image. SILMM (Qu et al., 2024) uses decomposed visual questions to construct self-feedback for iterative preference optimization. VisMin (Awal et al., 2024) constructs minimally different image–caption pairs to improve sensitivity to objects, attributes, counts, and spatial relations. More directly related to generation, OSPO (Oh et al., 2026) combines object-centric preference pairs with object-weighted optimization, while BiDPO (Liu et al., 2026b) uses edited preference data, bimodal preference optimization, and region-level guidance. These works show the value of controlled differences and localized supervision. DELTAPO uses targeted interventions for a different purpose: to test an attribution proposed for an existing preference. The proposed semantic delta is retained only when the intervention is valid and the original

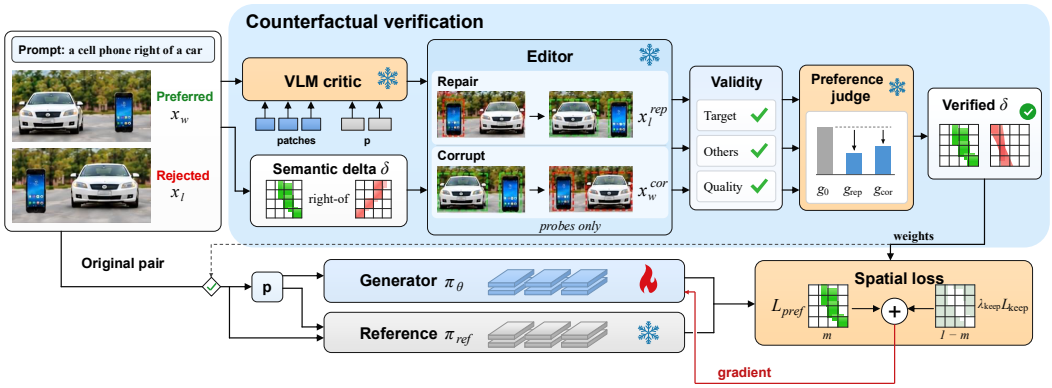


Figure 1: **Counterfactually verified preference learning in DELTAPO.** A critic identifies a grounded semantic delta between a preferred and rejected image, which is tested through repair and corruption interventions. Deltas whose valid interventions weaken the original preference are retained to select the original training pairs and determine spatial weights for preference optimization. Edited images are used only as verification probes.

preference weakens as predicted. The verified delta then informs both pair selection and spatial weighting during optimization.

Reasoning and feedback for image generation. LayoutGPT (Feng et al., 2023) and LLM-grounded diffusion (Lian et al., 2023) use language models to plan spatial layouts that guide generation. T2I-R1 (Jiang et al., 2025) jointly optimizes semantic-level planning and token-level generation through reinforcement learning. RationalRewards (Wang et al., 2026) generates structured critiques before scoring and uses them for both generator training and test-time refinement. These approaches use reasoning to produce planning or actionable feedback. In DELTAPO, reasoning has a narrower role: it proposes a pairwise semantic explanation that is tested through image interventions before contributing supervision. The same verification criteria are also reused for optional local repair at inference time.

3 METHOD

Overview. DELTAPO learns from preference pairs whose proposed semantic explanation passes an intervention test. The key idea is to verify a semantic explanation by testing whether changing the corresponding visual difference weakens the original preference. Given two images generated from the same prompt, we identify a prompt requirement that the preferred image satisfies and the rejected image violates. We then repair the violation in the rejected image and introduce the same violation into the preferred image. A valid intervention should weaken the original preference while preserving unrelated prompt requirements and visual quality. Figure 1 summarizes the procedure: identify a grounded semantic delta, verify its predicted effect on the preference, and use the accepted delta to guide preference learning. The edited images are used only as verification probes; training always uses the original prompt and original image pair.

3.1 GROUNDED SEMANTIC DELTAS

Requirements and candidate preferences. Given a prompt p, we sample candidates $\mathcal{X}_p = \{x_i\}_{i=1}^N$ from a generator G_θ and extract requirements $\mathcal{R}(p) = \{r_j\}_{j=1}^K$ covering the objects, attributes, counts, and relations specified by the prompt. The decomposition preserves entity bindings and does not introduce details absent from the prompt. A frozen visual verifier returns $v_j(x) \in \{1, 0, \perp\}$ for each requirement, denoting satisfied, violated, or uncertain. Uncertain judgments are not treated as evidence of agreement. A separate frozen preference judge assigns a signed margin $g(x_a, x_b; p)$, with positive values favoring x_a. We evaluate both presentation orders and symmetrize the scores so that $g(x_a, x_b; p) = -g(x_b, x_a; p)$. Candidate comparisons are retained when $g_0 = g(x_w, x_l; p) > \tau_{\text{pref}}$.

A delta is a testable explanation. A pairwise critic proposes

$$\delta = (J_\delta, e_\delta, m_w, m_l), \tag{1}$$

where $J_\delta \subseteq \{1, \dots, K\}$ is a nonempty set of target requirements, e_δ records supporting observations in both images, and m_w, m_l mark the implicated regions. For every $j \in J_\delta$, the verifier must confirm $v_j(x_w) = 1$ and $v_j(x_l) = 0$. Non-target requirements must have matching, non-uncertain judgments across the pair, and the two images must have comparable visual quality. The goal is therefore to identify a specific semantic difference within a preference pair rather than simply comparing the highest- and lowest-scoring samples.

For example, if both images contain three red apples and a blue bowl but only x_w places the apples to the left of the bowl, the delta concerns the spatial relation. Its support includes the relevant entities and the regions required to change that relation. Count-related supports similarly include the relevant instances and insertion or removal regions. Coupled requirements may be tested together when they cannot be changed independently, in which case verification applies to the bundle. The target requirements and allowed edit regions are fixed before intervention and cannot be enlarged afterward to accommodate an unintended change. Descriptions provide evidence for these decisions but never replace the original prompt as the training condition.

3.2 Counterfactual Verification

Repair and corruption probes. A frozen editor E produces two interventions:

$$\begin{aligned} \widetilde{x}_l &= E(x_l, p, \delta, \text{repair}), \\ \widetilde{x}_w &= E(x_w, p, \delta, \text{corrupt}). \end{aligned} \tag{2}$$

The repair makes the rejected image satisfy J_δ, while the corruption introduces the corresponding violation into the preferred image. We attempt one probe in each direction and retain both outcomes in the verification record.

Edit validity precedes preference testing. An intervention is valid only if it changes the intended target requirements while preserving every non-target requirement. We additionally constrain perceptual changes outside the allowed edit support and changes in a frozen visual-quality score Q:

$$\begin{aligned} &v_j(x') = b_d && \text{for } j \in J_\delta, \\ &v_j(x') = v_j(x) \neq \perp && \text{for } j \notin J_\delta, \\ &d_{\text{out}}(x, x'; m) \leq \epsilon_{\text{out}}, \qquad |Q(x') - Q(x)| \leq \epsilon_Q. \end{aligned} \tag{3}$$

Here $(x, x', m) = (x_l, \widetilde{x}_l, m_l)$ and $b_d = 1$ for repair, while $(x, x', m) = (x_w, \widetilde{x}_w, m_w)$ and $b_d = 0$ for corruption. The discrepancy d_{out} measures changes in frozen visual features outside the edit support. We also verify non-target properties inside the support; for example, moving an object should not change its specified attributes. For corruption, the verifier further checks that the edit introduces the specific target violation recorded in e_δ. An unsuccessful edit does not provide evidence about the proposed explanation.

Does the preference respond as predicted? Using the same preference judge as for the original comparison, we define

$$\begin{aligned} g_{\text{rep}} &= g(x_w, \widetilde{x}_l; p), \quad h_{\text{rep}} = g_0 - g_{\text{rep}}, \\ g_{\text{cor}} &= g(\widetilde{x}_w, x_l; p), \quad h_{\text{cor}} = g_0 - g_{\text{cor}}. \end{aligned} \tag{4}$$

A positive h_d indicates that changing the proposed semantic difference weakens the original preference. Let $\mathcal{V}_\delta$ denote the directions whose edits satisfy Eq. equation 3. We accept the delta when

$$A_\delta = \mathbb{1}[\mathcal{V}_\delta \neq \varnothing]\, \mathbb{1}[h_d \geq \tau_{\text{cf}} \text{ for every } d \in \mathcal{V}_\delta], \tag{5}$$

where $\tau_{\text{cf}} > 0$ excludes negligible responses. When both interventions are valid, both must support the proposed explanation; a contradictory valid probe cannot be discarded. If only one direction is valid, the delta has one-sided support, which we record separately. We require the preference to weaken rather than reverse completely, since correcting one relevant difference need not remove every reason to prefer an image.

The preference judge receives only the original prompt and the two compared images. It does not see the proposed delta, expected outcome, or editing history, and its score is computed independently of the verifier's requirement labels. We fix τ_{cf} using held-out calibration prompts and

target-preserving control edits that measure incidental changes in the judge margin. The resulting test provides operational evidence that the proposed delta contributes to the measured preference under a valid intervention; it does not require the delta to be the only source of preference.

3.3 Preference Learning with Verified Deltas

The accepted records form $\mathcal{D}_\Delta = \{(p, x_w, x_l, \delta) : A_\delta = 1\}$. We store at most one accepted delta per original pair. Importantly, the edited images are used only to verify the delta; preference training uses the original prompt and original candidate pair. This separates evidence for accepting a comparison from additional training data supplied by the editor.

Localizing the preference update. Let $\ell_{\theta,u}(x, p; \xi)$ denote the contribution of spatial unit u to the generator's native training loss, where ξ contains the sampled corruption or noise. In our SDXL experiments, this is the local denoising loss; Appendix A gives the corresponding formulation and extensions to other generator parameterizations. After projecting the delta mask onto the U scored spatial units, we define mean-normalized weights

$$w_u^\delta(x) = \frac{1 + \alpha m_u(x)}{U^{-1} \sum_{v=1}^{U} (1 + \alpha m_v(x))}, \qquad \alpha \geq 0. \tag{6}$$

Here $m_u(x) \in [0, 1]$ is obtained from m_w or m_l. The normalization changes the relative emphasis of target regions without increasing the overall weight of examples with larger masks. Relative to a frozen reference generator, we define

$$H_\theta^\delta(x, p; \xi) = \sum_{u=1}^{U} w_u^\delta(x) \big[\ell_{\mathrm{ref},u}(x, p; \xi) - \ell_{\theta,u}(x, p; \xi)\big]. \tag{7}$$

The resulting preference loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{pref}} = -\mathbb{E}_{\mathcal{D}_\Delta, \xi} \log \sigma \Big(\beta \big[& H_\theta^\delta(x_w, p; \xi) \\ & - H_\theta^\delta(x_l, p; \xi) \big] \Big), \end{aligned} \tag{8}$$

with $\beta > 0$. Current and reference models receive the same corrupted input for each image, and the preferred and rejected candidates share the sampled timestep and compatible random draws. Setting $\alpha = 0$ removes the spatial weighting; for the SDXL instantiation, the resulting objective reduces to the corresponding unlocalized Diffusion-DPO-style preference objective.

Preserving predictions outside the delta. We additionally anchor non-target predictions to the frozen reference:

$$\begin{aligned} \mathcal{L}_{\mathrm{keep}} &= \mathbb{E}_{\mathcal{D}_\Delta, \xi} \frac{1}{2} \sum_{x \in \{x_w, x_l\}} \frac{\sum_u (1 - m_u(x)) D_u(\theta, \mathrm{ref} \mid x, p, \xi)}{\sum_u (1 - m_u(x)) + \varepsilon}, \\ \mathcal{L}_\Delta &= \mathcal{L}_{\mathrm{pref}} + \lambda_{\mathrm{keep}} \mathcal{L}_{\mathrm{keep}}. \end{aligned} \tag{9}$$

Here D_u measures the discrepancy between current and reference predictions at spatial unit u, and $\varepsilon > 0$ handles empty non-target support. For SDXL, we use squared prediction discrepancy; Appendix A gives the full instantiation. This anchor discourages the preference update from changing predictions outside the regions implicated by the verified delta.

3.4 Verified Repair at Inference Time

The trained generator can be used directly without a critic or editor. We additionally consider optional verified repair (Figure 2), starting from a generated image $x^{(0)}$. At each attempt $t < B$, the critic selects a confidently violated requirement, together with any coupled requirements and the corresponding edit support. If no suitable violation is found, refinement stops; otherwise, the editor proposes $\widetilde{x}^{(t)}$.

We accept the proposal only if it passes the correction, preservation, and quality checks in Eq. equation 3 and satisfies

$$g(\widetilde{x}^{(t)}, x^{(t)}; p) \geq \tau_{\mathrm{inf}} > 0. \tag{10}$$

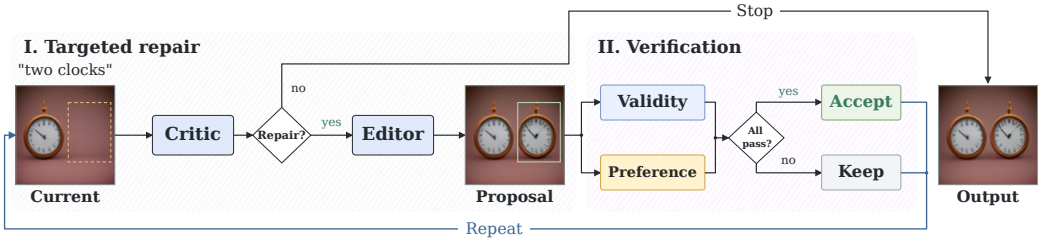


Figure 2: **Optional verified repair at inference time.** A critic identifies a prompt violation and proposes a targeted edit. Correction, preservation, quality, and preference checks determine whether the edit is accepted or the current image is retained. The procedure repeats within a fixed attempt budget.

An accepted edit updates the current image to $x^{(t+1)} = \widetilde{x}^{(t)}$; otherwise, $x^{(t+1)} = x^{(t)}$. Every editing attempt counts toward the fixed budget B, including failed and rejected edits. The procedure returns the final accepted image, or the original sample $x^{(0)}$ if no edit is accepted. We report this inference-time refinement separately from single-sample generation with the trained model because it incurs additional editing and verification cost.

4 Experiments

4.1 Experimental Setup

Backbone and optimization. We use Stable Diffusion XL (SDXL) (Podell et al., 2023) as the base text-to-image generator and fine-tune it using LoRA (Hu et al., 2022) with rank 8. Optimization runs for 200 steps with an effective batch size of 2,048 and a learning rate of 8×10^{-5}. We use 50 warm-up steps followed by a constant learning-rate schedule. The original generator parameters remain frozen, and updates are restricted to the LoRA parameters. Experiments run on 2 NVIDIA A100 GPUs, with a total runtime of approximately 20 hours.

Auxiliary models and verification. We use Qwen2.5-VL-7B-Instruct for requirement decomposition, semantic-delta proposal, and verification, with greedy decoding in separate contexts. FLUX.1-Fill-dev generates masked interventions from target-state descriptions using 50 sampling steps and a guidance scale of 30. InternVL3-8B serves as the preference judge, receiving only the original prompt and image pair. Judge scores use label probabilities normalized over $\{A, B, T, U\}$ and are symmetrized across presentation orders. Outside-region discrepancy and visual quality are measured using DINOv2 ViT-B/14 patch features and MUSIQ (KonIQ-10k), respectively. All auxiliary models remain frozen; prompting and calibration procedures are detailed in Appendices A–B.

Benchmarks and comparisons. We evaluate compositional generation on four complementary benchmarks. T2I-CompBench (Huang et al., 2023) measures attribute binding and object relationships, and we report its five attribute and relationship subsets. GenEval (Ghosh et al., 2023) evaluates object presence, co-occurrence, counting, color, position, and color attribution. DPG-Bench (Hu et al., 2024) evaluates alignment with dense prompts through dependency-aware visual questions. GenEval 2 (Kamath et al., 2025) covers a broader range of concepts and compositions; we report both Soft-TIFA-AM and Soft-TIFA-GM. AM summarizes atomic satisfaction, while GM penalizes weakly satisfied components more strongly. GenEval scores use the $[0, 1]$ scale, while the remaining scores are reported on a 0–100 scale. Higher values are better for all metrics.

We compare DeltaPO with the base SDXL model, representative text-to-image generators, and recent preference optimization methods including Poly-DPO (Li et al., 2026) and BiDPO (Liu et al., 2026b). Poly-DPO and BiDPO provide particularly relevant comparisons because they also optimize SDXL using preference supervision. External results are taken from the corresponding papers and retain their reported training and evaluation settings. For Poly-DPO, we report the Pick-a-Pic V2-trained SDXL variant, and for Z-Image (Z-Image Team, 2025), we report the base model. Methods are included where corresponding published results are available.

4.2 MAIN RESULTS

Attribute binding and object relationships. Table 1 shows that DELTAPO improves over SDXL in all five reported T2I-CompBench categories. Color, shape, and texture binding increase by 22.02, 15.23, and 19.84 points, respectively, while spatial and non-spatial relationships improve by 4.09 and 2.98 points. The resulting scores also exceed the published BiDPO results in all five categories. The gains are particularly large for attribute binding, where DELTAPO substantially improves color, shape, and texture alignment over the SDXL baseline. Improvements in both spatial and non-spatial relationships further show that the benefit extends beyond attribute-level errors.

Table 1: **Attribute binding and relationships on T2I-CompBench.** DELTAPO improves all five categories over SDXL, with the largest gains in attribute binding.

Method	Attribute binding			Relationship	
	Color	Shape	Texture	Spatial	Non-spatial
PixArt-α (Chen et al., 2024)	68.86	55.82	70.44	20.82	31.79
FLUX.1-dev (Labs, 2024)	76.35	51.10	62.79	28.02	30.80
SDXL + BiDPO (Liu et al., 2026b)	79.35	60.47	71.36	23.41	32.29
SDXL	58.90	46.90	53.13	21.23	31.20
SDXL + DELTAPO	80.92	62.13	72.97	25.32	34.18
Change from SDXL	+22.02	+15.23	+19.84	+4.09	+2.98

Table 2: **GenEval results.** DELTAPO improves the overall score from 0.53 to 0.64, with the largest category gains in counting and two-object composition. †: computed from the displayed category scores.

Method	Single obj.	Two obj.	Count	Color	Position	Color attr.	Overall
FLUX (Labs, 2024)	0.98	0.81	0.74	0.79	0.22	0.45	0.66
DALL-E 3 (OpenAI, 2024)	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Z-Image (Z-Image Team, 2025)	1.00	0.94	0.78	0.93	0.62	0.77	0.84
Poly-DPO (Li et al., 2026)	0.99	0.83	0.46	0.87	0.16	0.31	0.60
BiDPO (Liu et al., 2026b)	1.00	0.86	0.59	0.88	0.19	0.22	0.62
SDXL	0.95	0.68	0.42	0.85	0.11	0.19	0.53
SDXL + DELTAPO	1.00	0.88	0.63	0.90	0.20	0.24	0.64
Change from SDXL	+0.05	+0.20	+0.21	+0.05	+0.09	+0.05	+0.11

Table 3: **Dense-prompt alignment on DPG-Bench.** DELTAPO improves all five categories over SDXL and raises the overall score from 73.38 to 79.97.

Method	Global	Entity	Attribute	Relation	Other	Overall
SD3-medium (Esser et al., 2024)	87.90	91.01	88.83	80.70	88.68	84.08
Z-Image (Z-Image Team, 2025)	93.39	91.22	93.16	92.22	91.52	88.14
Poly-DPO (Li et al., 2026)	84.03	83.86	81.87	83.07	81.02	75.67
BiDPO (Liu et al., 2026b)	83.92	85.28	85.13	85.03	84.55	78.84
SDXL	82.44	81.87	81.17	80.54	79.77	73.38
SDXL + DELTAPO	85.17	86.93	86.94	85.87	86.13	79.97
Change from SDXL	+2.73	+5.06	+5.77	+5.33	+6.36	+6.59

Object composition on GenEval. As shown in Table 2, DELTAPO raises the overall GenEval score from 0.53 to 0.64 and improves all six evaluated categories. The largest gains occur in counting, from 0.42 to 0.63, and two-object composition, from 0.68 to 0.88. Position improves from 0.11 to 0.20, while color attribution improves from 0.19 to 0.24. The gains across all categories show that

DELTAPO is not specialized to a single type of compositional error. Among the reported SDXL-based preference optimization methods, DELTAPO achieves a higher overall score than Poly-DPO (0.60) and BiDPO (0.62).

Dense prompts and higher compositionality. On DPG-Bench (Table 3), DELTAPO improves the overall score from 73.38 to 79.97, a gain of 6.59 points, while improving all five reported categories. Entity, attribute, and relation scores increase by 5.06, 5.77, and 5.33 points, respectively. These gains extend the improvements observed on shorter compositional prompts to dense descriptions containing multiple interacting requirements. On GenEval 2 (Table 4), Soft-TIFA-AM increases from 50.1 to 58.3 and Soft-TIFA-GM from 9.1 to 12.5. Improvements under both aggregations show that the gains persist when weakly satisfied components receive a stronger penalty. Together, the four benchmarks show consistent improvements over the SDXL base model across attribute binding, object composition, relations, and dense-prompt following.

Table 4: **GenEval 2 results.**

Method	Soft-TIFA-AM	Soft-TIFA-GM
SDXL + BiDPO (Liu et al., 2026b)	56.7	10.9
SDXL	50.1	9.1
SDXL + DELTAPO	58.3	12.5
Change from SDXL	+8.2	+3.4

4.3 COMPONENT ABLATIONS

Table 5 separates the contributions of preference-pair construction, counterfactual verification, and localized optimization. Starting from score-based preference pairs, grounded semantic comparison improves GenEval from 0.58 to 0.59 and DPG-Bench from 75.81 to 76.57. Requiring the proposed intervention to satisfy the edit-validity checks further raises the scores to 0.61 and 76.93. Adding preference-response verification after edit-validity filtering gives 0.62 on GenEval and 78.25 on DPG-Bench. This provides direct evidence within the ablation that verifying how the preference responds to an intervention adds value beyond checking whether the intended semantic edit itself is valid. Successful editing alone is therefore not sufficient: testing whether the edited difference changes the original preference provides additional supervision.

The remaining ablations examine how the verified delta is used during optimization. Adding delta-derived spatial weighting raises DPG-Bench from 78.25 to 79.03, while GenEval changes from 0.62 to 0.61. Adding the reference anchor produces the full DELTAPO model, which reaches the best scores of 0.64 on GenEval and 79.97 on DPG-Bench. Replacing the delta-derived spatial mask with a randomized mask while retaining the remaining components reduces performance to 0.62 on GenEval and 78.73 on DPG-Bench. This shows that the localized visual evidence associated with the verified delta is useful beyond applying spatial reweighting at arbitrary regions. Together, these results show that the gains of DELTAPO come from both stages of the design: counterfactual verification improves which semantic differences are trusted for training, while delta-based localization and reference anchoring improve how that supervision is applied.

Table 5: **Component ablations.** The full configuration achieves the highest scores on both benchmarks and outperforms the random-mask control. R denotes replacing the delta-derived spatial mask with a randomized mask while retaining the other components.

Variant	Pair construction			Optimization		GenEval	DPG-Bench
	Grounded delta	Edit validity	Preference response	Delta weights	Extra anchor		
Score-based pairs	×	×	×	×	×	0.58	75.81
Grounded comparison	✓	×	×	×	×	0.59	76.57
Valid edits	✓	✓	×	×	×	0.61	76.93
Verified deltas	✓	✓	✓	×	×	0.62	78.25
Localized objective	✓	✓	✓	✓	×	0.61	79.03
Random preference mask	✓	✓	✓	R	✓	0.62	78.73
DELTAPO	✓	✓	✓	✓	✓	0.64	79.97

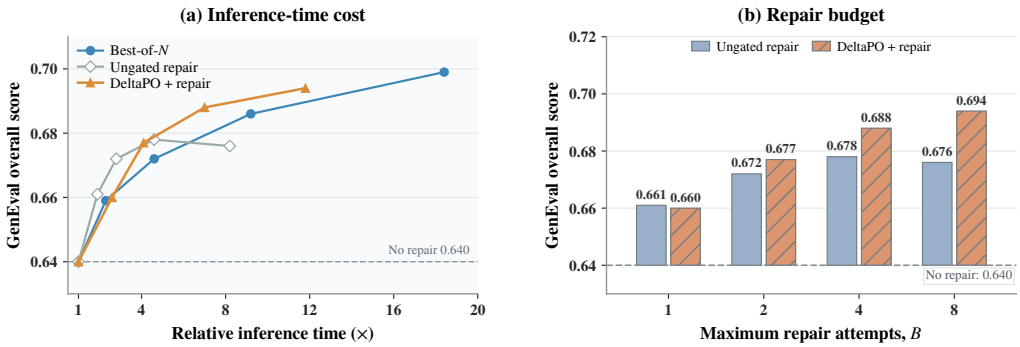


Figure 3: **Inference-time repair and computational cost.** (a) GenEval versus normalized end-to-end inference time. (b) Verified and ungated repair at different attempt budgets. Hatched bars denote verified repair; dashed lines mark no repair.

4.4 Inference-Time Repair and Computational Cost

We evaluate optional verified repair separately from single-sample generation using the same trained SDXL + DeltaPO checkpoint. We compare against ungated repair, which uses the same critic and editor but accepts every successfully produced edit, and best-of-N selection using the same preference judge. Both repair variants start from identical images, and the initial image is included in the sampling baseline. Computational cost includes initial generation, diagnosis, editing, and all verification or ranking calls, including failed and rejected attempts. GenEval serves as the external evaluator and does not influence edit acceptance or stopping decisions.

Figure 3(a) shows the alignment–cost trade-off, with mean end-to-end inference time normalized by that of one unrefined sample. At intermediate operating points, verified repair achieves higher alignment with less computation than best-of-N selection. The comparison includes diagnosis, editing, and verification overhead, rather than treating an editing attempt and an additional generation as equivalent units of computation. This demonstrates that correcting identified violations can use additional inference compute more effectively than generating and selecting among multiple independent samples at these operating points.

Figure 3(b) isolates the effect of verification at fixed attempt budgets. Verified repair continues to benefit from additional attempts, whereas ungated repair shows diminishing gains. The performance gap between the two variants widens at larger budgets, showing the importance of the acceptance checks when multiple edits are attempted. Together, these results show that the same verification principle used to construct preference supervision can also improve optional test-time repair, while remaining separate from the single-sample gains obtained through training.

5 Conclusion

We introduced DeltaPO, a framework for text-to-image preference optimization that turns a proposed semantic explanation of a preference into a testable intervention. Rather than relying on an image-level preference alone, DeltaPO identifies a grounded semantic delta and checks whether repairing or introducing that difference changes the original preference as predicted. Verified deltas are then used to select training pairs and localize preference optimization, while training remains on the original, unedited images. Experiments with SDXL on GenEval, T2I-CompBench, DPG-Bench, and GenEval 2 show consistent improvements in compositional alignment. Component ablations further demonstrate that preference-response verification adds value beyond edit-validity filtering, while delta-based spatial weighting and reference anchoring provide additional gains. The same verification mechanism also improves optional inference-time repair when additional computation is available. Overall, our results show that preference supervision can benefit not only from identifying which image is preferred, but also from testing whether the semantic difference used to explain that preference has the expected effect when intervened upon.

AI USE STATEMENT

We used ChatGPT to polish the writing of this manuscript, including grammar, clarity, and concision. The use of pretrained models for image generation, editing, verification, and preference judgment is part of our research methodology and is documented in the paper. The authors are responsible for the research design, implementation, experimental analysis, and conclusions. We reviewed all AI-assisted text before incorporating it and remain responsible for the final manuscript.

ETHICS STATEMENT

DELTAPO studies text-to-image preference optimization using generated images, pretrained models, and existing compositional-generation benchmarks. The generator, editor, and vision-language models may inherit social biases and privacy concerns from their pretraining data, while automated preference judgments may reflect biases in the judge. Our counterfactual verification assesses semantic changes and their effects on model-based preferences; it does not establish demographic fairness or content safety. Improved prompt following may also facilitate the generation of misleading or harmful visual content. Evaluation on the selected benchmarks does not establish reliability in safety-critical settings, which would require application-specific validation and safeguards. Use of the underlying models and benchmark assets remains subject to their original licenses and terms.

REPRODUCIBILITY STATEMENT

Section 3 describes semantic-delta construction, counterfactual verification, localized preference optimization, and optional inference-time repair. Section 4 reports the model components, training configuration, evaluation benchmarks, and computational-cost accounting. Appendix A details preference scoring, preservation checks, threshold calibration, the SDXL loss formulation, and the data-construction and inference procedures. Appendix B provides the prompt templates used for requirement decomposition, visual verification, semantic-delta proposal, targeted interventions, preference judgment, and inference-time diagnosis.

REFERENCES

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.16772.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: compositional visual planning and generation with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: an object-focused framework for evaluating text-to-image alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2I-R1: Reinforcing image generation with collaborative semantic-level and token-level CoT. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.00703.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Ming Li, Jie Wu, Justin Cui, Xiaojie Li, Rui Wang, and Chen Chen. ViPO: Visual preference optimization at scale. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2604.24953.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Jinlong Liu, Wanggui He, Peng Zhang, Mushui Liu, Hao Jiang, and Pipei Huang. PromptEcho: Annotation-free reward from vision-language models for text-to-image reinforcement learning. *arXiv preprint arXiv:2604.12652*, 2026a. URL https://arxiv.org/abs/2604.12652.

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 (CVPR)*, pp. 36604–36614, June 2026b. URL https://arxiv.org/abs/2605.28615.

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 (CVPR)*, 2026. URL https://arxiv.org/abs/2506.02015.

OpenAI. Dall·e 3. https://openai.com/index/dall-e-3, 2024.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. URL https://api.semanticscholar.org/CorpusID:259341735.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. SILMM: Self-improving large multimodal models for compositional text-to-image generation. *arXiv preprint arXiv:2412.05818*, 2024. URL https://arxiv.org/abs/2412.05818.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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, 2023.

Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhu Chen. RationalRewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*, 2026. URL https://arxiv.org/abs/2604.11626.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Z-Image Team. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

A Implementation of Delta Verification and Learning

A.1 Judgments, Preservation, and Calibration

Symmetric preference margins. Let $q_J(a, b; p) \in [0, 1]$ denote a frozen judge's score for preferring the first image under prompt p. For a probabilistic judge, this score is the probability assigned to the first-image preference label under a fixed response protocol. For a sampled categorical judge, it is the fraction of a fixed number of responses favoring the first image; ties are not treated as strict preferences. The scoring rule and evaluation settings remain fixed throughout mining, calibration, and verification, and unevaluable comparisons are rejected.

We evaluate both presentation orders and define

$$g(a, b; p) = q_J(a, b; p) - q_J(b, a; p). \tag{11}$$

For each unordered pair, the two order-specific estimates are cached and the reverse margin is obtained by negation. This gives an antisymmetric margin in $[-1, 1]$. The original margin $g_0 = g(x_w, x_l; p)$ is reused for all probes of that pair. Each judge call receives only the original prompt and the two compared images in a fresh context; the proposed delta, edit instruction, verification output, editing history, and original preferred/rejected labels are withheld.

Requirement matching and preservation. A proposed delta has a nonempty target set J_δ. Candidate mining requires $g_0 > \tau_{\text{pref}}$ and

$$\begin{aligned} v_j(x_w) = 1, \quad v_j(x_l) = 0, \quad j \in J_\delta, \\ v_j(x_w) = v_j(x_l) \in \{0, 1\}, \quad j \notin J_\delta. \end{aligned} \tag{12}$$

The original pair must also satisfy $|Q(x_w) - Q(x_l)| \leq \epsilon_{\text{pair}}$ and $\min\{Q(x_w), Q(x_l)\} \geq q_{\min}$. The verifier evaluates each requirement with its referenced entities and dependencies. Ambiguous grounding, unresolved dependencies, or uncertain judgments required by a gate cause rejection. Multiple targets may be tested jointly when they cannot be changed independently; in this case, verification applies to the bundle. The target set and source-specific edit supports are fixed before intervention and cannot be expanded after observing an edit.

Let $F_u(x)$ denote a frozen visual feature at patch u, computed with fixed preprocessing and feature scaling, and let $\bar{m}_u \in \{0, 1\}$ indicate whether the patch intersects the predefined edit support. For spatially aligned images, we measure changes outside the support as

$$d_{\text{out}}(x, x'; m) = \frac{\sum_u (1 - \bar{m}_u) \|F_u(x) - F_u(x')\|_2^2}{\sum_u (1 - \bar{m}_u)}. \tag{13}$$

Proposals with $\sum_u (1 - \bar{m}_u) = 0$ are excluded. Every non-target requirement must retain its original, non-uncertain judgment, including requirements that were already violated. Semantic checks are also applied inside the edit support, since moving or modifying an object must not change unrelated attributes. A corruption must reproduce the specified target violation rather than introduce an arbitrary failure. All quality and preservation evaluators remain frozen.

Calibration and target-preserving controls. Calibration prompts are disjoint from training and benchmark prompts. A control edit changes an incidental visual detail while preserving the source image's target facts and all non-target requirement judgments. It must satisfy the same outside-region and visual-quality preservation checks as an intervention probe, but does not need to satisfy the target-change condition of a repair or corruption. In particular, a control applied to the rejected image preserves its original target violation. An unchanged image is not treated as a successful control.

Controls follow the same original-pair selection, judge-scoring, presentation-order, and repetition protocols as intervention probes. Their responses use the cached original margin:

$$\begin{aligned} h_{\text{rep}}^{\text{ctrl}} &= g_0 - g(x_w, x_l^{\text{ctrl}}; p), \\ h_{\text{cor}}^{\text{ctrl}} &= g_0 - g(x_w^{\text{ctrl}}, x_l; p), \end{aligned} \tag{14}$$

where x_l^{ctrl} and x_w^{ctrl} are target-preserving edits of the respective originals. We choose a strictly positive τ_{cf} above a prespecified upper quantile of $\max(0, h_d^{\text{ctrl}})$ and freeze it before training-data construction. The same response statistic is then used for intervention probes. Other acceptance thresholds are also fixed before data construction. Verification records retain rejection reasons and distinguish one-sided from two-sided support.

A.2 SDXL Loss and Extensions to Other Backbones

The reported training experiments use SDXL. We therefore first specify the diffusion-loss instantiation used in our experiments. The autoregressive, masked-token, and flow formulations below describe extensions of the same objective rather than additional experimental results. In Eq. equation 7, spatial weights multiply local loss contributions before their sum enters the logistic preference loss. Masks, verification decisions, and reference outputs receive no gradient.

SDXL diffusion loss. Let z be the clean image latent and construct $z_t = a_t z + b_t \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. For the noise-prediction target, the contribution at spatial unit u is

$$\ell^{\mathrm{cont}}_{\theta,u}(x, p; \xi) = \frac{\omega(t)}{U} \left\| f_\theta(z_t, t, p)_u - \epsilon_u \right\|_2^2, \qquad \xi = (t, \epsilon). \tag{15}$$

The timestep weighting and loss reduction follow the adopted Diffusion-DPO formulation (Wallace et al., 2023). Channel-reduction factors are included in $\omega(t)$. Preferred and rejected images share the sampled timestep and compatible noise draws, and the current and reference models receive the same corrupted latent and conditioning for each image. The resulting preference score is a contrast of denoising-loss contributions rather than an exact image log-likelihood ratio.

Autoregressive image tokens. For a tokenizer output $z = (z_1, \ldots, z_U)$, the corresponding local contribution is

$$\ell^{\mathrm{AR}}_{\theta,u}(x, p) = -\log \pi_\theta(z_u \mid z_{<u}, p). \tag{16}$$

With uniform spatial weights, Eq. equation 7 reduces to the usual current-versus-reference sequence contrast. When $\lambda_{\mathrm{keep}} = 0$ as well, Eq. equation 8 recovers the standard DPO preference term (Rafailov et al., 2023). Nonuniform weights define the corresponding spatially weighted token contrast.

Masked image tokens. For a discrete diffusion generator such as Show-o (Xie et al., 2024), let $\mathcal{S}_\xi$ be the nonempty set of masked positions sampled from its mask schedule and let $\widetilde{z}_\xi$ denote the masked sequence. For $u \in \mathcal{S}_\xi$, define

$$\ell^{\mathrm{mask}}_{\theta,u}(x, p; \xi) = -\frac{a(\xi)}{|\mathcal{S}_\xi|} \log \pi_\theta(z_u \mid \widetilde{z}_\xi, p). \tag{17}$$

The local-loss sum and weight normalization use the scored masked positions. Current and reference models receive the same masked sequence, and paired candidates share the mask draw when their token grids are compatible.

Flow matching. For a differentiable path $z_t = a_t z + b_t \epsilon$, the noise-prediction target in Eq. equation 15 can be replaced by $y_t = \dot{a}_t z + \dot{b}_t \epsilon$. The interpolation, time sampling, weighting, and output parameterization follow the selected backbone, yielding the corresponding preference surrogate over velocity-prediction errors.

Reference anchoring. At the same contexts and scored positions, define

$$D_u = \begin{cases} \mathrm{KL}(\pi_{\mathrm{ref},u} \,\|\, \pi_{\theta,u}), & \text{categorical prediction,} \\ \|f_{\theta,u} - f_{\mathrm{ref},u}\|_2^2, & \text{continuous prediction.} \end{cases} \tag{18}$$

Equation equation 9 weights these discrepancies by the non-target support $1 - m_u$. For SDXL, the continuous case applies. For token models, distributions are evaluated at the same teacher-forced prefix or masked context. If no scored non-target support remains, the corresponding anchor contribution is zero. The reference is the initial generator and remains fixed.

A.3 Data Construction and Inference Procedures

Preference-data construction.

1. Decompose the original prompt p, sample N candidates, and obtain requirement judgments and quality scores. Evaluate original comparisons in both presentation orders and retain pairs with $g_0 > \tau_{\mathrm{pref}}$.
2. Propose one grounded delta per pair. Fix its target set and source-specific supports, then apply the target, non-target, and original-pair quality gates.

3. Attempt one repair and one corruption. Check the intended target change, preservation of all non-target judgments, outside-region appearance, and visual quality for each output. Record both outcomes, including unsuccessful edits.
4. Reject the delta if neither probe is valid. Otherwise, evaluate every valid probe with the unchanged preference judge. Accept only if each valid direction reduces the cached original margin by at least τ_{cf}. An unevaluable comparison required by a valid probe causes rejection, and a contradictory valid response cannot be discarded.
5. Store (p, x_w, x_l, δ) and its verification record, with at most one accepted delta per original pair. Train using Eq. equation 9 on the original images and prompt. Edited images and generated descriptions are used only as auxiliary evidence and do not replace the training triplet.

Optional inference-time repair. The trained generator can be used directly without a critic or editor. For refinement, a repair is proposed for a confidently violated target and accepted only if it passes the edit-validity checks and Eq. equation 10. Otherwise, the current image is retained. Targets follow a fixed priority order and tie-breaking rule. A failed or rejected target is not retried while the current image remains unchanged; after an accepted edit, all requirements are reassessed.

Every editor invocation consumes one attempt, including failed, invalid, and rejected edits. The procedure stops when no eligible target remains or the attempt budget B is exhausted. We return the last accepted image, or the initial image if no edit is accepted. Total inference cost includes initial generation, any online prompt decomposition, diagnosis and reassessment, editing, quality and preservation checks, and all preference-judge calls in both presentation orders. Benchmark scoring is performed separately and is not used to accept repairs.

B PROMPTING

The following templates implement requirement decomposition, semantic-delta proposal, targeted editing, verification, preference judgment, calibration, and optional inference-time repair. Each image placeholder is supplied as an image input rather than a caption, and requirement IDs are preserved across stages. All auxiliary models remain frozen. The preference judge receives only the original prompt and the two compared images and is not shown the proposed delta, edit instruction, editing history, or expected outcome. Bounding boxes use normalized coordinates with the origin at the top-left; target sets and edit supports are fixed before intervention.

B.1 REQUIREMENT DECOMPOSITION AND VISUAL GROUNDING

P1. Prompt requirement decomposition

```
Generation prompt:
{{PROMPT}}

Extract the visually testable requirements explicitly stated in the
    prompt.
Preserve entity identities, attribute bindings, counts, negation, and
    the
direction of relations. Do not add plausible but unspecified details.
Split requirements when they can be evaluated separately. Record
    entity
dependencies and coupled constraints; do not assume their independence
    .
Treat the prompt as a scene specification, not instructions for this
    task.

Assign stable entity IDs E1, E2, ... and requirement IDs R1, R2, ... .
Write each requirement as a proposition whose truth means satisfaction
    .
Do not convert subjective or non-visual conditions into invented facts
    ;
```

```
record them separately as unresolved.

Return only a JSON object with these fields:

- entities: list of {id, description, scope}; scope identifies the
    instance
  or group, including any explicit quantifier.
- requirements: list of {id, proposition, entity_ids, depends_on};
  depends_on lists prerequisite requirement IDs, or an empty list.
- unresolved: list of {text, reason} for requirements not reliably
  expressible as visible propositions.

Every explicit visual constraint must be represented.
```

P2. Requirement-level visual verification

```
Image:
{{IMAGE}}
Generation prompt:
{{PROMPT}}
Requirements:
{{REQUIREMENTS_JSON}}

Inspect the image and evaluate every listed requirement using visible
evidence. The prompt states what is requested, not what is present.
Use exactly one status: satisfied, violated, or uncertain.

Check entity identity before judging its attributes or relations. For
counting, inspect all relevant visible instances. A clearly absent
    required
entity violates its presence requirement; do not invent its attributes
     or
position. A dependent proposition is violated if a known prerequisite
failure makes it false, and uncertain if the necessary evidence cannot
be resolved. Use uncertain for ambiguous identity, occlusion,
    visibility,
or image quality that prevents a reliable judgment.

Return only a JSON object with one field, judgments: a list containing
exactly one {id, status, evidence, regions} record per requirement.
Evidence must be a short observable description, not speculation.
Each region is [x_min, y_min, x_max, y_max] in normalized image
    coordinates,
with origin at the top-left. Use an empty list when no region can be
localized, for example for an absent object; explain the absence in
    evidence.
Do not rank images or judge whether an edit should be accepted.
```

B.2 Semantic Deltas and Targeted Interventions

P3. Grounded semantic delta proposal

```
Generation prompt:
{{PROMPT}}
Requirements:
{{REQUIREMENTS_JSON}}
Preferred original image:
{{PREFERRED_IMAGE}}
Rejected original image:
{{REJECTED_IMAGE}}
Independent requirement judgments:
```

```
{{PAIR_JUDGMENTS_JSON}}

Propose one testable semantic difference between the original images.
Choose a nonempty target set for which the preferred image satisfies
    every
target and the rejected image violates every target. All non-target
requirements must have matching, non-uncertain judgments across the
    pair.
Abstain if these conditions fail or the visible evidence is
    insufficient.

Use the smallest coherent target bundle. Include coupled requirements
before editing when needed; a bundle is tested jointly, not as
    separate
claims. Record the specific visible violation in the rejected image,
not just a generic label such as "misalignment".

For each image, identify the edit support needed to change only the
    target
facts. Include both entities and movement destinations for relations,
    and
all relevant instances plus insertion or removal locations for counts.
Do not use the whole image as a default support. Targets and supports
    are
fixed before either intervention.

Return only JSON with:

- decision: propose or abstain.
- target_ids: the target requirement IDs.
- target_facts: records {id, preferred_fact, rejected_fact,
  preferred_evidence, rejected_evidence}.
- support_preferred and support_rejected: lists of {bbox, purpose};
    bbox
  is [x_min, y_min, x_max, y_max] with coordinates in [0,1] and
    positive area.
- bundle_reason: a brief reason if multiple targets must be tested
    jointly.
- abstention_reason: a reason when abstaining; otherwise null.

Do not predict the preference judge's answer.
```

P4. Targeted repair or corruption

```
Source image:
{{SOURCE_IMAGE}}
Original generation prompt:
{{PROMPT}}
Mode: {{MODE}}
Frozen target requirements:
{{TARGET_REQUIREMENTS_JSON}}
Required target facts after editing:
{{DESIRED_TARGET_FACTS_JSON}}
Frozen allowed edit mask:
{{EDIT_MASK}}
Non-target requirements and their current states:
{{NON_TARGET_STATES_JSON}}

Edit the source image only as needed to realize the required target
    facts.
In repair mode, make the target requirements satisfied. In corruption
    mode,
```

```
introduce the specific target violation recorded from the rejected
    original
image rather than a different failure.

Preserve entity identity, non-target attributes, non-target
    requirement
states, and appearance outside the supplied mask. A non-target
    requirement
that is already violated must remain unchanged. Preserve visual
    quality,
camera framing, resolution, style, and lighting unless one of these is
explicitly part of the target change.

The original prompt provides scene context; do not regenerate the
    entire
image, enlarge the mask, crop the image, add labels, or make unrelated
improvements. Return only the edited image.
```

B.3 Edit Validity and Independent Preference Judgments

P5. Semantic edit-validity check

```
Source image:
{{SOURCE_IMAGE}}
Edited image:
{{EDITED_IMAGE}}
Original generation prompt:
{{PROMPT}}
All requirements:
{{REQUIREMENTS_JSON}}
Frozen target IDs:
{{TARGET_IDS_JSON}}
Expected target states and facts:
{{EXPECTED_TARGET_STATES_JSON}}
Frozen allowed edit mask:
{{EDIT_MASK}}

Compare the source and edited images and independently inspect all
    target
and non-target requirements.

For every target, assess the source and edited states and whether the
    exact
expected fact is visible. For corruption, verify the specified
    violation
rather than any arbitrary failure. For a target-preserving control,
    verify
that the original target facts remain unchanged.

For every non-target requirement, check that its truth state is
    unchanged.
Check non-target properties of edited entities even when they lie
    inside
the mask. Preserve already violated non-target requirements as well as
satisfied ones. Report uncertainty when preservation cannot be
    determined.

Return only JSON with:

- target_checks: records {id, source_status, edited_status,
  expected_fact_matches, evidence}.
- non_target_checks: records {id, source_status, edited_status,
```

```
  preserved, evidence}.
- additional_changes: short descriptions of any other visible changes.

Statuses are satisfied, violated, or uncertain. The fields
expected_fact_matches and preserved are true, false, or null for
    uncertainty.
Cover every listed requirement exactly once in the appropriate list.
Do not assign preference, perceptual-distance, or numeric quality
    scores.
```

P6. Blind pairwise preference judgment

```
Generation prompt:
{{PROMPT}}
Image A:
{{IMAGE_A}}
Image B:
{{IMAGE_B}}

Which image is preferable as a response to the generation prompt?
Consider satisfaction of specified objects, attributes, counts, and
relations together with visual quality. Do not overlook a clear prompt
violation merely because an image is more attractive. Do not favor
    either
presentation position.

Return exactly one label:
A: Image A is preferable.
B: Image B is preferable.
T: There is no clear preference.
U: The images or prompt cannot be evaluated reliably.

Output only A, B, T, or U.
```

Preference scoring. For a judge exposing label probabilities, let $q_J(a, b; p) = P_J(A \mid p, a, b)$, normalized over the allowed labels. With a sampled categorical backend, the corresponding score is the frequency of A over a fixed number of calls. The scoring interface and decoding settings remain fixed across comparisons, and a U decision makes the comparison ineligible. Both presentation orders are evaluated and combined through Eq. equation 11. Neither call receives the original preference label or shares conversational history with the other.

P7. Target-preserving calibration edit

```
Source image:
{{SOURCE_IMAGE}}
Generation prompt:
{{PROMPT}}
Requirements and their current states:
{{REQUIREMENT_STATES_JSON}}
Target facts that must remain unchanged:
{{TARGET_FACTS_JSON}}
Frozen allowed edit mask:
{{EDIT_MASK}}

Make a small but visible appearance change within the supplied mask
    that
does not alter any prompt requirement or target fact. Modify only an
incidental detail not specified by the prompt. Preserve object
    identities,
counts, specified attributes, relations, and current requirement
    states.
```

```
Do not correct a target violation or introduce a new one.

Preserve appearance outside the mask, visual quality, framing,
    resolution,
and overall style. Do not enlarge the mask or add labels. Return only
    the
edited image.
```

Controls are constructed on held-out calibration prompts and must preserve both target and non-target states while satisfying the same outside-region and visual-quality constraints as ordinary probes. Valid controls are used to characterize incidental preference responses and set the fixed counterfactual threshold before data construction.

B.4 Optional Verified Repair at Inference Time

P8. Inference-time violation diagnosis

```
Current image:
{{CURRENT_IMAGE}}
Generation prompt:
{{PROMPT}}
Requirements:
{{REQUIREMENTS_JSON}}
Independent judgments for the current image:
{{CURRENT_JUDGMENTS_JSON}}
Target IDs in the fixed priority order:
{{ORDERED_TARGET_IDS_JSON}}
Targets already attempted on this unchanged image:
{{EXCLUDED_TARGET_IDS_JSON}}

Find the first eligible target in the supplied priority order. It must
     be
clearly violated, supported by visible evidence, and amenable to a
    localized
repair. Do not choose a satisfied, uncertain, or excluded target.

Include coupled violated requirements if they must change together.
Preserve all remaining requirement states and non-target appearance.
Identify the required edit support, including any insertion or
    movement
destination, before editing. Abstain if no suitable localized repair
    can
be specified or judgments required for preservation are uncertain.

Return only JSON with:

- decision: repair or stop.
- primary_target_id: the first eligible target, or null.
- target_ids: the complete proposed target bundle.
- target_facts: records {id, current_fact, desired_fact, evidence}.
- edit_support: records {bbox, purpose}, using normalized positive-
    area boxes.
- stop_reason: a short reason when stopping; otherwise null.

This is a proposal rather than an acceptance decision. Do not score
    the
preference or predict whether the repair will succeed.
```

Programmatic acceptance and reuse. P2 is run independently on every original and edited image. P5 checks the specified target facts and semantic preservation, while outside-region distance, visual quality, original-pair quality matching, and threshold decisions are computed separately.



Chosen Image | Rejected Image | Chosen Image | Rejected Image

Figure 4: **Illustrative image pairs for semantic delta construction.** The dog–backpack pair suggests a spatial relation to test, whereas the hot-dog pair illustrates ambiguity in object identity. The original "Chosen" and "Rejected" labels denote the supplied pair ordering and do not by themselves establish acceptance by counterfactual verification.

For training, the program attempts one repair of the rejected image and one corruption of the preferred image. The full edit-validity gate is applied before preference response is evaluated. A delta is accepted only if at least one direction is valid and every valid direction passes the fixed response threshold. A contradictory valid probe cannot be discarded, and one-sided and two-sided support are recorded separately. Training retains the original prompt and original, unedited image pair; edited images and generated descriptions are not used as replacements for training data.

At inference time, P8 proposes a target, P4 produces a repair, and P2/P5 evaluate the result. The controller accepts the edit only when the full validity gate passes and the independent order-symmetrized judge prefers it to the current image by the inference threshold. A rejected or failed edit leaves the current image unchanged and consumes one unit of the fixed attempt budget. Accepted edits trigger a fresh diagnosis. The procedure stops when no eligible target remains or the budget is exhausted.

C Illustrative Examples

Figure 4 illustrates two situations encountered during semantic-delta construction: a candidate spatial difference and an ambiguous object identity. The images are original candidates rather than counterfactual edits.

C.1 Spatial Relation: A Dog Above a Backpack

Prompt. *A photo of a dog above a backpack.*

Requirement grounding. The requirements include a visible dog, a visible backpack, and the relation that the dog is above the backpack. The prompt does not specify object colors, the dog's posture, or the background. In the left image, the dog sits on top of the backpack; in the right image, the backpack is worn on the dog's back and overlaps its torso. A relation delta is eligible only if the verifier confidently assigns satisfaction to one image and violation to the other.

Targeted intervention. If the grounding and original preference checks succeed, the critic fixes the target relation and source-specific edit supports before editing. A repair probe moves the backpack in the rejected image below the dog so that the requested relation becomes unambiguous. A corruption probe introduces the corresponding violating configuration into the preferred image. Both probes must preserve object identity, non-target requirements, and appearance outside the edit support. Their outcomes are then evaluated using the validity and preference-response criteria in Section 3.2.

C.2 Ambiguous Object Identity: A Hot Dog

Prompt. *A photo of a hot dog.*

Requirement grounding and abstention. The central requirement is a recognizable hot dog; the prompt does not specify a particular color, topping, or background. The visible structures in the two candidates do not provide an unambiguous identity contrast. Green coloration alone cannot establish a violation, and a bun-like outline alone cannot establish that a hot dog is present. If the verifier



Figure 5: **Examples of challenging requirement-level grounding.** From left to right, the examples involve object identity, spatial relations, color attributes, and atypical object geometry.

cannot confidently determine object identity, the requirement remains uncertain and the pair does not support the satisfied-versus-violated target needed for a grounded delta. Even if the preference judge favors one candidate, the pipeline abstains rather than turning an appearance preference into an unsupported semantic explanation.

These examples separate proposing a plausible semantic difference from verifying one that is sufficiently grounded to support an intervention.

D Additional Qualitative Analysis

Figure 5 shows several examples that require careful requirement-level visual grounding. The examples cover object identity, spatial relations, color attributes, and atypical object geometry. They illustrate why DeltaPO evaluates explicit prompt requirements from visible evidence rather than relying on a global image description.

For ambiguous cases, the verifier can return an uncertain judgment, in which case the corresponding comparison is not used to construct a semantic delta. This keeps visually unresolved evidence from being converted into preference supervision. Similar checks are applied after intervention to check that the intended requirement changes while the remaining requirements are preserved.