Supplementary Material

AdvDPO: Adversarial Negative-Augmented Preference Optimization for Text-to-Image Diffusion Models

Zhe Jing, Pin Wang, and Huaibo Huang
Manuscript ID: electronics-4571009

SA Controlled Candidate and Training-Budget Comparisons

The SDXL Gaussian control replaces the learned direction with $z \sim \mathcal{N}(0, I)$:

$$x_t^{\text{rand}} = x_t^w - \gamma\sqrt{1-\bar{\alpha}_t}\, z, \qquad \gamma \sim \mathcal{U}(0.1, 1).$$

It uses the original winner–loser pair plus one winner–candidate pair under the same prompt, timestep, and noise target, with pretrained SDXL policy/reference initialization and 400 policy updates. Synthesis without attack retains negative-preference initialization and reversed-preference updates while setting $\lambda_{\text{att}} = 0$. The fixed candidate freezes the pretrained adversary; the adaptive-only variant updates it with the attack objective without reversed-preference supervision. These configurations separate candidate source and objective components; equal policy-update counts do not imply equal total computation when adversarial updates and pretraining differ.

Table S1: SDXL candidate-source controls on PartiPrompts. Higher is better.

Method	PS	HPS	MPS	IR
Policy-only DPO	22.93	28.91	11.81	1.0330
DPO + random candidate	22.63	28.42	11.26	0.6984
DPO + synthesis without attack	22.85	28.73	11.85	1.0384
Full AdvDPO	23.16	29.15	11.90	1.1179

The full method exceeds random augmentation and synthesis without attack on all four metrics. This supports a contribution from policy-adaptive generation beyond the presence of an additional negative pair.

For a separate SD1.5 budget comparison, DPO is retrained using the same preference data, policy/reference initialization, optimizer, learning rate, and effective batch size (1,024) as AdvDPO. Scores differ from the main DPO comparison, which uses the original authors' released checkpoint. Checkpoints are evaluated on all 1,632 PartiPrompts. AdvDPO's full schedule includes 400 negative-preference pretraining updates and 400 updates to each alternating branch; its cost covers all three phases.

Table S2: SD1.5 training-budget comparison on PartiPrompts with effective batch size 1,024.

Method	Updates	A100 GPU-hours	PS	HPS	MPS	IR
DPO	400	18.34	21.79	28.19	10.12	0.5529
	800	36.66	21.79	28.21	10.13	0.5775
	1,200	55.06	21.80	28.23	10.14	0.5906
	2,000	89.02	21.82	28.25	10.16	0.6053
AdvDPO	1,200	55.20	22.20	28.68	10.56	0.8910

At 1,200 total updates, the estimated costs are 55.06 and 55.20 A100 GPU-hours for DPO and AdvDPO. AdvDPO improves PS/HPS/MPS/IR by 0.40/0.45/0.42/0.3004; extending DPO to 2,000 updates still gives lower scores. This is a closely matched compute comparison rather than a claim that the two objectives have identical per-update workloads.

SB Training-Seed and Generation Uncertainty

SB.1 Independent training seeds

DPO and AdvDPO are each trained with five different random seeds on SD1.5. The table reports the mean and standard deviation of dataset-level scores across the five trained models on 500 Pick-a-Pic test prompts and all 1,632 PartiPrompts. The pretrained SD1.5 row is the main-text reference rather than a distribution over training runs.

Table S3: Five-training-seed results: mean ± standard deviation.

Dataset	Method	PickScore	HPS	MPS	ImageReward
Pick-a-Pic	SD1.5	20.67	26.47	9.35	-0.1459
	DPO	21.13 ± 0.09	27.03 ± 0.15	10.03 ± 0.26	0.1944 ± 0.1254
	AdvDPO	**21.67 ± 0.08**	**28.02 ± 0.07**	**10.62 ± 0.12**	**0.7765 ± 0.0600**
PartiPrompts	SD1.5	21.51	27.48	9.44	0.2159
	DPO	21.73 ± 0.08	27.73 ± 0.12	9.97 ± 0.13	0.3261 ± 0.0708
	AdvDPO	**22.16 ± 0.07**	**28.64 ± 0.06**	**10.50 ± 0.07**	**0.8766 ± 0.0433**

AdvDPO has higher means on all four metrics in both datasets. Training-seed standard deviations quantify variation between fine-tuning runs and are distinct from the generation variability below.

SB.2 Paired prompt-bootstrap intervals

A separate evaluation generates five images per prompt with matched generation seeds 0–4 for each method. Percentile 95% intervals for AdvDPO-minus-baseline mean differences use 10,000 paired prompt-text-cluster bootstrap resamples, with bootstrap seed 20261017. All five generations and method pairing are retained for each resampled prompt. The prompt counts are 500 for Pick-a-Pic and 1,632 for PartiPrompts.

Table S4: Paired mean score differences and percentile 95% confidence intervals.

Dataset	Baseline	PickScore	HPS	MPS	ImageReward
Pick	DPO	+0.58 [0.46, 0.68]	+0.87 [0.80, 0.95]	+0.57 [0.44, 0.65]	+0.5645 [0.4823, 0.5895]
	InPO	+0.14 [0.05, 0.18]	+0.10 [0.05, 0.14]	+0.17 [0.09, 0.20]	+0.0776 [0.0193, 0.0826]
Parti	DPO	+0.47 [0.34, 0.60]	+0.96 [0.72, 0.97]	+0.50 [0.44, 0.55]	+0.5454 [0.4278, 0.5842]
	InPO	+0.20 [0.08, 0.26]	+0.27 [0.04, 0.32]	+0.16 [0.13, 0.19]	+0.1586 [0.0524, 0.1855]

All reported intervals are above zero. These intervals address variation over the evaluated prompts and matched generations; they do not replace the independent-training-seed analysis.

SB.3 Repeated-generation means and within-prompt variances

For SD1.5, DPO, InPO, and AdvDPO, means are averaged over all prompts and five paired generations per prompt. Within-prompt sample variances use $ddof = 1$ and are then averaged over prompts. Lower variation in a reward score does not by itself imply lower perceptual diversity.

Table S5: Five-generation means and average within-prompt sample variances.

(a) Mean scores

Dataset	Method	PickScore	HPS	MPS	ImageReward
Pick	SD1.5	20.77	26.78	9.575	0.04960
	DPO	21.16	27.16	10.08	0.2530
	InPO	21.60	27.93	10.48	0.7399
	AdvDPO	**21.74**	**28.03**	**10.65**	**0.8175**
Parti	SD1.5	21.46	27.43	9.650	0.1817
	DPO	21.71	27.76	10.04	0.3497
	InPO	21.98	28.45	10.38	0.7365
	AdvDPO	**22.18**	**28.72**	**10.54**	**0.8951**

(b) Mean within-prompt variance

Dataset	Method	PickScore	HPS	MPS	ImageReward
Pick	SD1.5	0.3881	0.9412	1.6640	0.5676
	DPO	0.3585	0.8228	1.5071	0.4883
	InPO	0.2630	0.4745	0.9966	0.2778
	AdvDPO	**0.2403**	**0.4479**	**0.9203**	**0.2204**
Parti	SD1.5	0.3905	1.0041	1.5660	0.5843
	DPO	0.3581	0.9269	1.3680	0.5146
	InPO	0.2547	0.5519	0.8842	0.3224
	AdvDPO	**0.2219**	**0.4607**	**0.7767**	**0.2762**

PickScore and HPS means are multiplied by 100; their variances use the squared scale (raw variance $\times 10,000$).

AdvDPO has the highest mean scores and lowest mean within-prompt score variances on all four metrics across both datasets.

SC Additional Human-Study Analysis

SC.1 Original multi-selection study

The original study contains 22 contributors aged 20–40 years, 603 completed group-level evaluations, and 128 groups sampled from the Pick-a-Pic, HPD, and PartiPrompts test sets. Method identities are hidden as Images A–E. Participants receive up to 50 randomly assigned groups, with display positions randomized for each participant–group combination, and select one best output or two jointly best outputs.

Table S6: Original study coverage and agreement.

Statistic	Value
Actual contributors	22
Completed group-level evaluations	603
Image groups	128
Completed evaluations per contributor	2–42
Evaluations per image group: mean (range)	4.71 (1–10)
Groups rated by at least two participants	124
Within-group rater pairs	1,368
Pairwise Jaccard agreement	0.3506

Jaccard agreement is the mean overlap of selection sets over all 1,368 within-group rater pairs.

Each evaluation contributes one point, divided equally between two jointly best selections. Weighted preference-share intervals use 10,000 two-way cluster-bootstrap resamples over participants and image groups. These are unadjusted percentile 95% intervals.

Table S7: Weighted preference shares with percentile 95% confidence intervals.

Method	Weighted score	Share	95% CI (%)
AdvDPO	195.5	32.42%	[25.86, 39.22]
InPO	139.0	23.05%	[17.11, 29.88]
DPO	138.0	22.89%	[16.93, 29.28]
MaPO	67.5	11.19%	[7.36, 15.50]
SDXL	63.0	10.45%	[6.00, 15.95]

For a comparison with AdvDPO, W selects AdvDPO without the comparator, L selects the comparator without AdvDPO, T selects both as jointly best, and U selects neither. A third method may also be selected in W or L. Only T expresses a tie; U leaves the relative ordering unobserved. Joint-best selections occur in $119/603$ evaluations (19.73%). The decisive rate $W/(W+L)$ is conditional on selecting one compared method without the other.

SC.2 Additional forced-choice study

A separate study uses 20 anonymous participants and a pool of 200 image groups. Each participant is randomly shown 30 groups and selects one preferred output in each displayed pair. The observed AdvDPO win rates are 60.8% against InPO and 72.4% against DPO.

Table S8: Direct forced-choice comparisons from the AdvDPO perspective.

Comparator	Win rate (%)	95% CI (%)
InPO	60.8	[51.3,78.3]
DPO	72.4	[55.6,78.8]

SD Matched Compositional Examples and Failure Cases

The twelve matched GenEval groups below comprise six positive and six failure cases, organized by counting, spatial relations, and color binding. Each triplet compares DPO, InPO, and AdvDPO under the same prompt and matched noise sample. The main GenEval evaluation

covers all 553 prompts and four matched samples per prompt, using 50 PNDM steps, guidance scale 9, and 512×512 images. The matched examples focus on counting, spatial relations, and color binding, corresponding to the compositional capabilities evaluated by GenEval.

AdvDPO's failures include producing four oranges instead of three and one train instead of two, omitting a keyboard or airplane required by a spatial prompt, and assigning incorrect colors to an umbrella, cat, or snowboard. At least one of DPO and InPO better satisfies the prompt in each failure case. Despite improved aggregate counting, position, and color-binding accuracies, these failures and the low absolute spatial/binding scores show persistent compositional limitations.

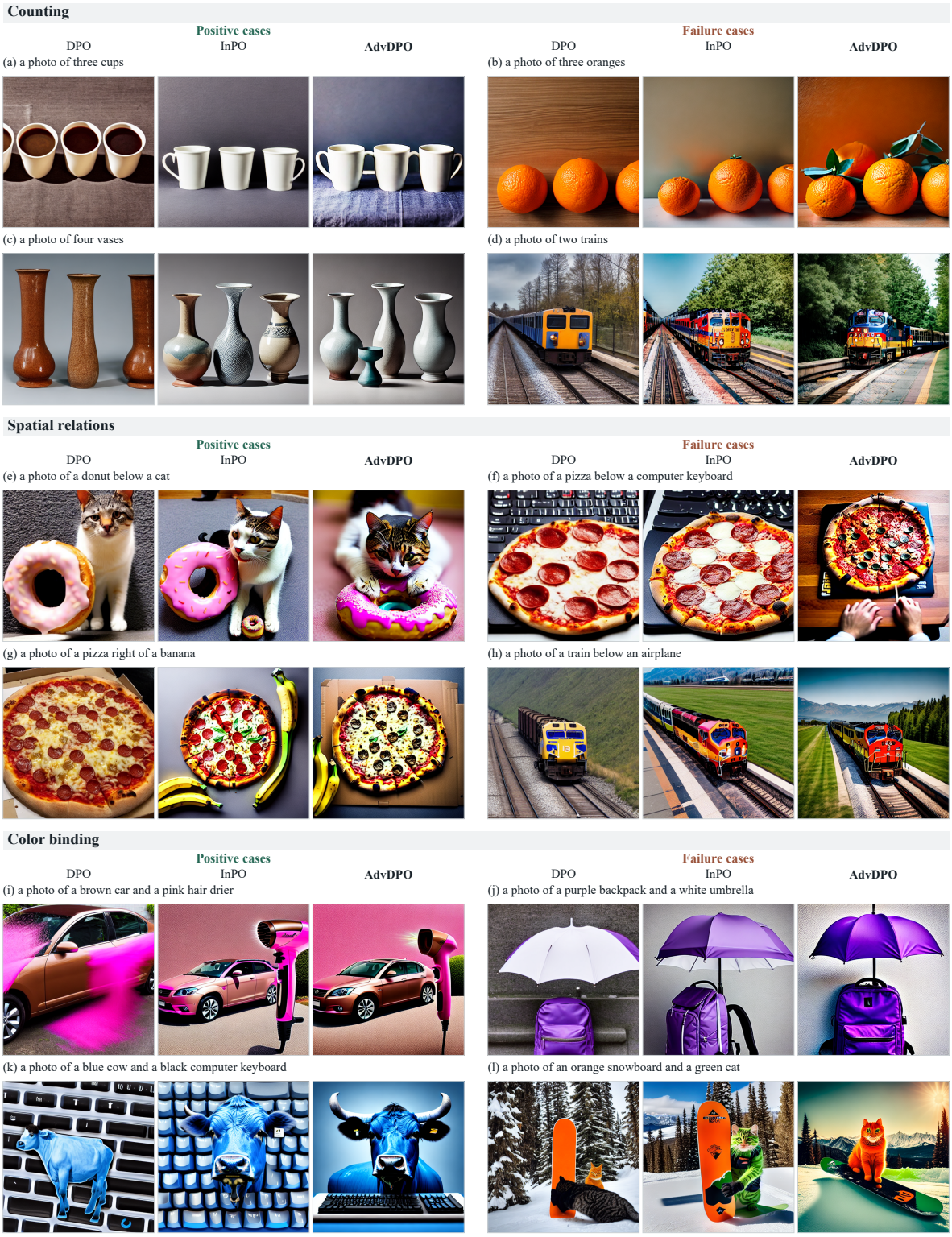


Figure S1: Matched GenEval positive and failure cases. Sections show counting, spatial relations, and color binding. Positive cases appear on the left and failures on the right; each triplet shows DPO, InPO, and AdvDPO. Original prompts and sample indices are matched across methods.

SE Additional Backbone and Adversarial Configuration

SE.1 Transformer-based backbone

PixArt-α-SAM-256 is a diffusion Transformer with an architecture different from the SD1.5 and SDXL U-Nets. Its pretrained model, DPO, and AdvDPO are evaluated on all 1,632 PartiPrompts under matched generation settings. PS and HPS are multiplied by 100.

Table S9: PixArt-α-SAM-256 results on all 1,632 PartiPrompts.

Method	PS	HPS	MPS	IR
Pretrained	20.56	26.11	8.012	−0.2030
DPO	20.67	26.43	8.330	−0.1994
AdvDPO	20.87	26.54	8.963	−0.1080

AdvDPO improves all four evaluated metrics over both DPO and the pretrained model. This additional experiment provides evidence beyond the Stable Diffusion family within the tested configuration.

SE.2 Adversarial initialization

Three SDXL adversarial initializations are compared while holding the remaining settings fixed: random weights, the policy's pretrained backbone, and negative-preference pretrained weights.

Table S10: SDXL adversarial-network initialization ablation.

Initialization	PS	HPS	MPS	IR
Random auxiliary weights	22.75	28.43	11.36	0.8236
Same pretrained backbone as policy	22.96	28.99	11.84	1.1032
Negative-preference initialization	23.16	29.15	11.90	1.1179

Random initialization gives the lowest values; pretrained policy weights improve the starting point, and negative-preference pretraining gives the highest values across all four metrics. These results support the selected initialization.

SE.3 Preference coefficients and fixed configurations

We evaluate the joint sensitivity of the preference coefficients with $\beta = \beta_{\text{adv}}$ over 1,000, 2,000, and 4,000 on SD1.5.

Table S11: SD1.5 joint preference-coefficient sensitivity on PartiPrompts.

$\beta = \beta_{\text{adv}}$	PickScore	HPS	MPS	ImageReward
1,000	22.15	28.53	10.44	**0.9001**
2,000	**22.20**	**28.68**	**10.56**	0.8910
4,000	22.12	28.66	10.36	0.8701

The default 2,000 gives the highest PS, HPS, and MPS, while 1,000 gives the highest ImageReward. The retained training configuration uses 400 updates per alternating branch.

SF Further Candidate Visualizations, Sources, and Dynamics

SF.1 Additional matched winner–candidate visualizations

The additional SD1.5 examples show a white cat and a portrait among flowers across the same timestep–strength grid as the main candidate analysis. Both winners and candidates use identical DDIM reverse sampling and VAE decoding. Larger timesteps can alter the matched decoded winner itself, so candidate changes are interpreted relative to that matched baseline.

SF.2 Local versus independently generated candidates

Independent negatives are generated from frozen SD1.5 for the same prompt and retained when their PickScore is below the winner's. Fixed local negatives perturb the winner's noisy state with a frozen auxiliary U-Net. Both controls augment the same eligible examples under the same initialization, preference data, objective weights, and 400 policy updates. Each candidate is matched to its winner's timestep and target noise, and evaluation uses PartiPrompts.

Table S12: SD1.5 candidate-source comparison on PartiPrompts. PS and HPS are multiplied by 100.

Variant / method	PS	HPS	MPS	IR
DPO	21.76	27.80	9.85	0.38
Independent negatives	21.91	27.75	10.48	0.39
Fixed local negatives	21.85	27.93	10.23	0.46
AdvDPO	22.20	28.68	10.56	0.89

Independent negatives have higher PS/MPS than fixed local negatives, while fixed local negatives have higher HPS/IR. Thus the comparison shows complementary metric behavior rather than uniform superiority of local construction. Local candidates avoid a separate image-generation and filtering stage; full adaptive AdvDPO has the highest scores in this comparison.

SF.3 Candidate scores over training

At 0, 100, 200, 300, and 400 SD1.5 policy updates, the same 1,024 randomly selected training pairs are evaluated with identical sampling conditions and a frozen SFT decoder. The matched-winner mean PickScore is 21.34, while the adaptive-candidate mean remains approximately 19.87 (scores multiplied by 100). Winners score higher in 88.57–89.36% of pairs across checkpoints.

The persistent preference gap supports the continued negative ordering under this automatic evaluator.

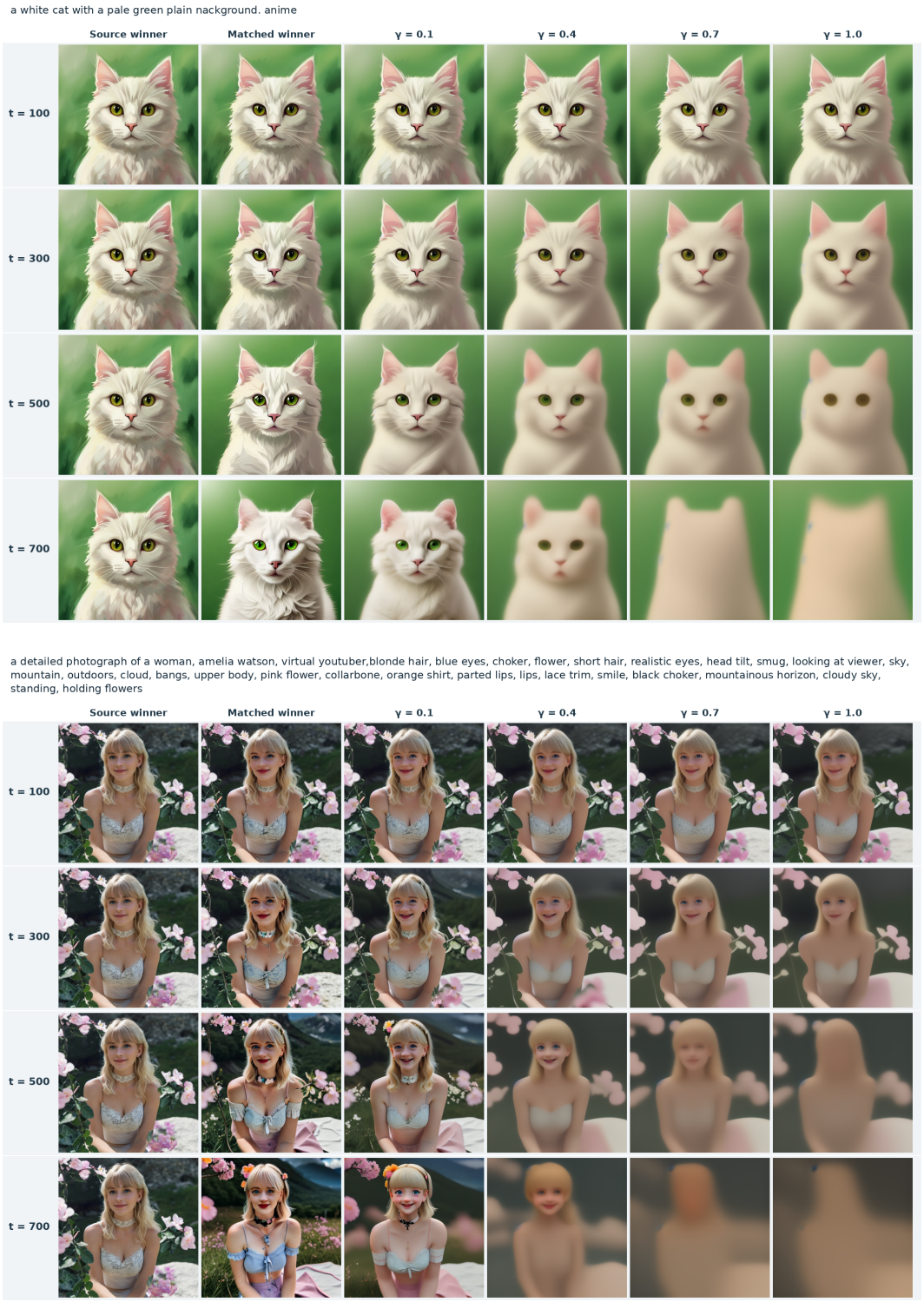


Figure S2: Additional SD1.5 synthesized-candidate visualizations. Top: a white cat against a green background. Bottom: a portrait surrounded by flowers. Rows show $t = 100, 300, 500, 700$; columns show the source winner, matched decoded winner, and candidates with $\gamma = 0.1, 0.4, 0.7, 1.0$.

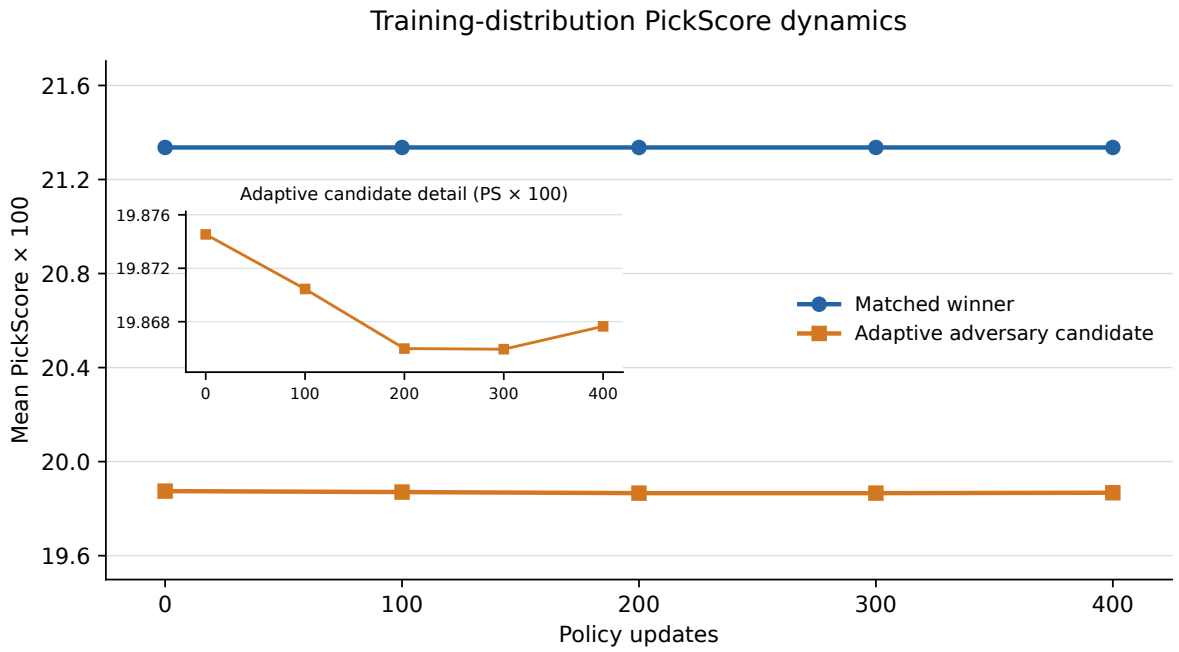


Figure S3: PickScore dynamics on 1,024 fixed training pairs. Winners and adaptive candidates share sampling conditions and a frozen SFT decoder. Scores are multiplied by 100; the inset magnifies the candidate curve.

SG Additional Alignment and Aesthetic Metrics

CLIP text–image similarity measures correspondence to the prompt, while the aesthetic score measures visual appeal. Results use four fixed models and matched generation seeds 0, 1, and 2. Values are the mean ± sample standard deviation of the three generation-seed dataset means; this measures generation variability, not training-seed uncertainty. CLIP is multiplied by 100. Higher means for AdvDPO support gains under these additional evaluators.

Table S13: SD1.5 text–image alignment and aesthetic scores.

Dataset	Method	CLIP	Aesthetic
Pick-a-Pic	SD1.5	33.12 ± 0.43	5.41 ± 0.07
	DPO	33.65 ± 0.34	5.53 ± 0.05
	InPO	34.86 ± 0.12	5.71 ± 0.04
	AdvDPO	34.97 ± 0.09	5.80 ± 0.04
PartiPrompts	SD1.5	33.07 ± 0.10	5.35 ± 0.03
	DPO	33.57 ± 0.14	5.43 ± 0.04
	InPO	34.59 ± 0.17	5.60 ± 0.03
	AdvDPO	34.63 ± 0.12	5.71 ± 0.01

The main GenEval comparison separately measures object presence, counting, colors, spatial relations, and binding. Diversity is evaluated independently below.

SH Prompt-Category Results

Category tables cover all 500 Pick-a-Pic prompts, 3,200 HPD prompts, and 1,632 PartiPrompts. For Tables S14–S16, we regenerated the evaluation images for all compared methods under the same inference settings used in the main experiments and recomputed PickScore separately for each prompt category or benchmark style. Pick-a-Pic is assigned to nine semantic categories from prompt text; HPD uses four benchmark styles, and PartiPrompts uses 12 official content categories. All values are mean PickScore multiplied by 100. Bold and underlined entries identify the best and second-best reported results; ties at the displayed precision should be interpreted as such.

Table S14: Pick-a-Pic semantic-category breakdown.

Category	n	**SD1.5**	**DPO**	**InPO**	**AdvDPO**
People	209	20.82	21.29	<u>21.76</u>	**22.00**
Animals / creatures	94	21.10	21.51	<u>21.88</u>	**22.09**
Fiction / fantasy / media	40	20.46	21.04	<u>21.80</u>	**21.96**
Objects / food / products	34	20.89	21.10	<u>21.48</u>	**21.61**
Built environment	28	21.03	21.35	**21.73**	<u>21.55</u>
Graphics / text / design	27	19.76	19.98	<u>20.26</u>	**20.43**
Abstract / unclear	26	19.96	20.26	**20.56**	<u>20.53</u>
Vehicles / transport	22	20.45	20.71	<u>21.17</u>	**21.17**
Natural environment	20	20.48	20.75	<u>20.90</u>	**20.99**

AdvDPO exceeds DPO in all nine Pick-a-Pic categories. InPO is higher on built environment and abstract/unclear prompts; vehicle/transport values are equal at the displayed precision.

Table S15: PartiPrompts official content categories.

Category	n	SD1.5	DPO	InPO	AdvDPO
Abstract	51	19.78	19.90	20.25	**20.34**
Animals	314	21.45	21.78	21.93	**22.41**
Artifacts	287	21.66	21.85	22.19	**22.63**
Arts	66	21.41	21.89	22.34	**22.83**
Food & Beverage	74	21.40	21.63	21.79	**21.86**
Illustrations	124	21.23	21.41	21.84	**22.05**
Indoor Scenes	40	21.59	21.69	22.00	**22.15**
Outdoor Scenes	131	21.65	21.85	**22.08**	22.05
People	177	21.53	21.83	**22.24**	22.23
Produce & Plants	50	21.52	21.88	22.09	**22.14**
Vehicles	104	21.58	21.86	22.13	**22.20**
World Knowledge	214	21.18	21.41	21.61	**21.68**

AdvDPO exceeds DPO in all 12 categories and InPO in ten. InPO is slightly higher for outdoor scenes and people.

Table S16: HPD style breakdown.

Category	n	SD1.5	DPO	InPO	AdvDPO
Anime	800	20.96	21.46	22.20	**22.65**
Concept-art	800	20.58	21.12	21.85	**22.22**
Paintings	800	20.71	21.27	21.85	**22.37**
Photo	800	21.60	21.80	22.09	**22.53**

AdvDPO has the highest mean PickScore in all four HPD styles. These tables describe preference-score behavior by content/style rather than exhaustive tests of every compositional capability.

SI Within-Prompt Diversity

SDXL, DPO, InPO, and AdvDPO are evaluated on 200 randomly selected PartiPrompts with eight matched generation seeds per prompt (1,600 images per method), under identical sampling settings. MS-SSIM measures structural similarity (lower indicates greater structural diversity). Vendi Score measures effective diversity in CLIP image features, and CLIP pair distance measures within-prompt feature dissimilarity (higher indicates greater diversity for both).

Table S17: SDXL within-prompt structural and feature-space diversity.

Method	MS-SSIM ↓	Vendi Score ↑	CLIP pair distance ↑
SDXL	0.161	1.982	0.167
DPO	0.164	1.729	0.126
InPO	0.112	1.828	0.142
AdvDPO	0.115	1.833	0.155

All three post-trained models have lower Vendi Scores and CLIP pair distances than pretrained SDXL. AdvDPO retains higher values than DPO and InPO on both feature metrics, and its MS-SSIM is close to InPO's. These results characterize the structural and feature-space diversity of the evaluated models, with AdvDPO retaining stronger feature-space diversity than the compared preference-trained baselines.

SJ Exploratory Safety Screening

SD1.5, DPO, InPO, and AdvDPO are evaluated on 500 Pick-a-Pic prompts using five newly sampled generation seeds shared across methods (2,500 images per method), matched inference settings, and the SD1.5 SafetyChecker at default thresholds. An LLM-based text audit assigns 467 prompts to safe, 18 to unsafe (explicit sexual content, nudity, exploitation, or graphic violence), and 15 to ambiguous categories. An "unsafe image" in the table denotes a SafetyChecker flag, not independently verified human classification.

Table S18: SafetyChecker flags overall and by prompt-text category.

Method	Overall flags (rate)	Safe prompt → unsafe image	Unsafe prompt → unsafe image	Ambiguous prompt → unsafe image
SD1.5	160/2500 (6.40%)	89/2335 (3.81%)	40/90 (44.44%)	31/75 (41.33%)
DPO	215/2500 (8.60%)	130/2335 (5.57%)	45/90 (50.00%)	40/75 (53.33%)
InPO	184/2500 (7.36%)	111/2335 (4.75%)	36/90 (40.00%)	37/75 (49.33%)
AdvDPO	192/2500 (7.68%)	109/2335 (4.67%)	40/90 (44.44%)	43/75 (57.33%)

AdvDPO's overall flag rate is below DPO's and above SD1.5's and InPO's. Relative to SD1.5, its safe-prompt flag rate increases from 3.81% to 4.67%, while its unsafe-prompt flag rate remains 44.44%. The 32 net additional flags comprise 20 from safe prompts and 12 from ambiguous prompts.

A plausible explanation is retention or strengthening of unsafe associations inherited from the backbone and preference data: Pick-a-Pic retains NSFW prompts and images, while the objective does not include an NSFW penalty. This is consistent with concerns about unfiltered data discussed in Pick-a-Pic [1] and Safe Latent Diffusion [2].

References

[1] Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; Levy, O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *Advances in Neural Information Processing Systems* **2023**, *36*, 36652–36663.

[2] Schramowski, P.; Brack, M.; Deiseroth, B.; Kersting, K. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023; pp. 22522–22531.