ALIGNED, BUT MORE CAPABLE? EVALUATING POST-TRAINED TEXT-TO-IMAGE MODELS BEYOND PREFERENCE METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference alignment is widely used to improve text-to-image generation by making model outputs better match human preferences. Yet it remains unclear whether such gains reflect broader generation improvements or primarily a redistribution of probability toward preferred outputs. In this work, we systematically study eight representative preference-alignment methods from three complementary perspectives: preference improvement, broader generation capability, and responsible behavior. Using Best@k evaluation under increasing sampling budgets, we first confirm that alignment makes human-preferred outputs easier to sample. Beyond preference-oriented evaluation, however, these gains often diminish as the sampling budget increases, with the base model matching or surpassing aligned models on several benchmarks. Prompt-level analysis provides further insight: alignment does not uniformly improve success across prompts, but instead shifts the success distribution, increasing repeated-generation success for some prompts while leaving more prompts unsolved within the same budget. These changes also extend beyond capability, as most evaluated methods increase unsafe-generation rates and all increase occupational gender imbalance relative to the base model. Finally, we show that agentic generation with a frozen base model can achieve competitive generation performance while yielding a lower average unsafe-generation ratio. Overall, our results suggest that preference alignment reshapes the generation distribution toward preferred outputs, but these benefits do not consistently extend to broader generation quality or responsible behavior. This motivates assessing not only preference gains, but also how they transfer to broader capabilities and affect other model behaviors.

1 INTRODUCTION

Preference alignment has become an important stage in text-to-image post-training, with the goal of steering pretrained generators toward outputs that better reflect human preferences. Recent methods leverage supervision from human preference annotations (Wallace et al., 2024; Liu et al., 2026) or curated high-quality images (Chen et al., 2026b), and have reported substantial improvements over their pretrained base models on standard generation benchmarks (Ghosh et al., 2023; Hu et al., 2024) and learned preference metrics (Kirstain et al., 2023). These advances suggest that preference alignment can increase the likelihood of generating outputs favored by the alignment signal.

However, stronger performance on preference-oriented evaluations does not necessarily imply that the model's generation capabilities have improved more broadly. The same benchmark gain may reflect more reliable success on prompts already solved by the base model, successful generations on prompts for which the base model produces no successful samples within the evaluated budget, or both. Distinguishing these effects is important for understanding what preference alignment actually changes. Moreover, reshaping the generation distribution toward a particular preference signal may also affect behaviors outside the optimization objective. Recent work has documented reward over-optimization and distributional changes during preference optimization (Zhang et al., 2024; Chen et al., 2026a), yet comparatively little is known about how such changes relate to broader generation capability, safety, and demographic behavior.

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.

This motivates a broader question: *what changes when a pretrained text-to-image model is aligned toward human preference?* We study this from three perspectives: whether preference gains extend to broader capability, how success is distributed across prompts, and whether alignment affects responsible behavior such as unsafe generation and occupational gender bias. We further ask whether some improvements can be obtained without modifying the generator itself.

To this end, we evaluate eight representative preference-alignment methods (Wallace et al., 2024; Liu et al., 2026; Chen et al., 2026b) spanning diverse optimization paradigms and supervision sources, from human preference data to curated high-quality images. For generation capability, we use Best@k evaluation under increasing sampling budgets across six benchmarks covering compositional generation (Ghosh et al., 2023), dense prompt following (Hu et al., 2024), world knowledge (Niu et al., 2026), human preference (Saharia et al., 2022), creative generation (Li et al., 2026a), and image realness (Leng et al., 2026). We further analyze prompt-level success distributions and solvable-prompt coverage to characterize where alignment gains arise. For responsibility, we evaluate unsafe generation (Qu et al., 2023) and occupational gender bias (Cho et al., 2023). Finally, we investigate agentic image generation (He et al., 2026), where an external multimodal agent refines and selects generations while keeping the underlying generator fixed, as an alternative way to exploit the capabilities of the pretrained model.

Across these settings, we find that preference alignment has more selective effects than standard benchmark scores alone suggest. Our main findings are:

- **Preference gains do not consistently translate into broader capability gains.** Preference alignment makes preferred outputs easier to sample, but on several broader benchmarks its advantage decreases as the sampling budget grows, with the base model matching or surpassing aligned models at larger k (Section 4.1).
- **Alignment reshapes how generation success is distributed across prompts.** On WISE, aligned models do not exhibit a uniform shift toward higher prompt-level success. Instead, more prompts appear in both the high-success and zero-success regimes, and every evaluated method gains fewer newly solvable prompts than it loses previously solvable ones under $k = 32$ (Section 4.2).
- **The effects of alignment extend beyond generation quality.** Most evaluated methods increase unsafe-generation rates, while all increase occupational gender imbalance relative to the base model. Meanwhile, agentic generation with a frozen base model achieves competitive generation performance while yielding a lower average unsafe-generation ratio, suggesting an alternative route to exploiting pretrained generation capability without directly reshaping the generator distribution (Sections 5 and 6.1).

2 Related Work

Preference alignment for text-to-image generative models. Preference alignment has become a common strategy for adapting pretrained text-to-image generators toward human-preferred outputs. Existing methods broadly differ in how preference signals are obtained and how they are optimized. Early work applies reinforcement learning directly to diffusion models, including DDPO (Black et al., 2024) and DPOK (Fan et al., 2023), while differentiable-reward approaches such as AlignProp (Prabhudesai et al., 2023) and DRaFT (Clark et al., 2024) backpropagate reward signals through the generation process. Diffusion-DPO (Wallace et al., 2024) instead directly optimizes pairwise preferences without an explicit reward model. More recent methods extend preference optimization to flow-based generators, including Flow-GRPO (Liu et al., 2026), DanceGRPO (Xue et al., 2025), DiffusionNFT (Zheng et al., 2026), and GRPO-Guard (Wang et al., 2026). Complementary to these optimization-centric approaches, RealAlign (Chen et al., 2026b) constructs preference supervision from curated high-quality images and degraded counterparts, showing that useful alignment signals need not come solely from explicit human preference labels.

Evaluation and analysis of preference alignment. Recent work has increasingly examined the limitations of preference alignment beyond standard benchmark improvements. Several studies report reward over-optimization or reward hacking, where gains on proxy rewards do not necessarily translate into better generation quality (Zhang et al., 2024; Hong et al., 2026). Related work further shows that aggressive preference optimization can reduce generative diversity and concentrate outputs around narrow high-reward modes (Jena et al., 2025; Chen et al., 2026a). Other analyses

identify objective and estimation mismatches in preference optimization that can distort the learned generation distribution (Zhu et al., 2025; Wu et al., 2025).

While prior work has primarily focused on understanding and mitigating optimization failures in preference alignment, such as over-optimization and mode collapse, our work instead examines its broader empirical effects on reachable generation capability and responsible behavior.

3 Experiment Setup

Evaluated preference alignment methods. We consider two categories of alignment methods based on their supervision signals: human preference data and curated high-quality data. For the former, we evaluate Flow-GRPO (Liu et al., 2026), GRPO-Guard (Wang et al., 2026), Diffusion-NFT (Zheng et al., 2026), Diffusion-DPO (Wallace et al., 2024), Flow-OPD (Fang et al., 2026) and Diffusion-OPD (Li et al., 2026b). For the latter, we use the framework by Chen et al. (2026b) with two different curated data sources: the original RealAlign dataset and a CivitAI dataset (wallstoneai, 2025), denoted as RealAlign and CivitaiAlign, respectively. All methods use `Stable Diffusion 3.5 Medium` (SD-3.5-M) (Esser et al., 2024) as the common base model, enabling a controlled comparison across different preference alignment approaches.

Evaluation Benchmarks. For *generation capability*, we evaluate six complementary aspects of text-to-image generation: compositional generation using GenEval (Ghosh et al., 2023), dense prompt following using DPG-Bench (Hu et al., 2024), world knowledge using WISE (Niu et al., 2026), creative generation using Creative split of Qwen-Image-Bench (Li et al., 2026a), and human preference using DrawBench (Saharia et al., 2022). We additionally evaluate image realness (Leng et al., 2026) on prompts from MSCOCO (Lin et al., 2014).

For *responsibility*, our evaluation focuses on unsafe generation and occupational gender bias. Safety is assessed on the `Lexica`, `4chan`, and `Template` prompt sets introduced in Unsafe Diffusion (Qu et al., 2023). Occupational gender bias is evaluated following DALL-Eval (Cho et al., 2023), which measures gender imbalance in generations conditioned on different occupations.

Evaluation Metrics. For *generation capability*, we retain each benchmark's original evaluator and scoring criteria while extending evaluation to multiple sampling budgets. Standard evaluation summarizes average performance across generated samples. To examine whether alignment gains persist when models receive more generation attempts, we evaluate Best@k for increasing values of k. For a given prompt, Best@k is the expected maximum evaluator score among k independently sampled generations, with results averaged across prompts. For binary-valued metrics, this reduces to pass@k, the probability that at least one of the k generations succeeds (Chen et al., 2021). For continuous-valued metrics, it measures the expected highest score among the samples. Unless otherwise specified, we evaluate sampling budgets up to $k = 32$. We estimate Best@k using unbiased estimators for binary-valued (Chen et al., 2021) and continuous-valued metrics (Nakano et al., 2021).

For *responsibility*, we report the unsafe-generation rate and occupational gender imbalance. For each unsafe prompt, we sample multiple generations and compute the unsafe-generation rate as the proportion of samples classified as unsafe; we then average this rate across prompts. For occupational gender bias, we compute $p_{\text{male}} = \frac{n_{\text{male}}}{(n_{\text{male}}+n_{\text{female}})}$ and $p_{\text{female}} = 1-p_{\text{male}}$ for each occupation, where n_{male} and n_{female} count male- and female-classified images. We classify occupations as male-biased if $p_{\text{male}} \geq 0.6$, female-biased if $p_{\text{male}} \leq 0.4$, and neutral otherwise. We also report mean absolute deviation (MAD), computed for each occupation as $\text{MAD}_{\text{gender}} = \frac{1}{2}(|p_{\text{male}} - 0.5| + |p_{\text{female}} - 0.5|)$ and averaged across all 83 occupations. Unsafe-generation rate and MAD range from 0 to 1 and 0 to 0.5, respectively, with lower values indicating safer generation and greater gender balance. Detailed evaluation protocols are provided in Appendix A.

4 Preference Alignment's Effect on Generation Capability

We investigate whether the gains from preference alignment reflect higher reachable generation capability or primarily improved sampling efficiency. We first examine the overall trend using Best@k under increasing sampling budgets, where $k = 32$ in our experiments. We then analyze the per-prompt success distribution to understand how alignment changes the likelihood of generating successful outputs. Finally, we compare solvable-prompt coverage between the base and aligned

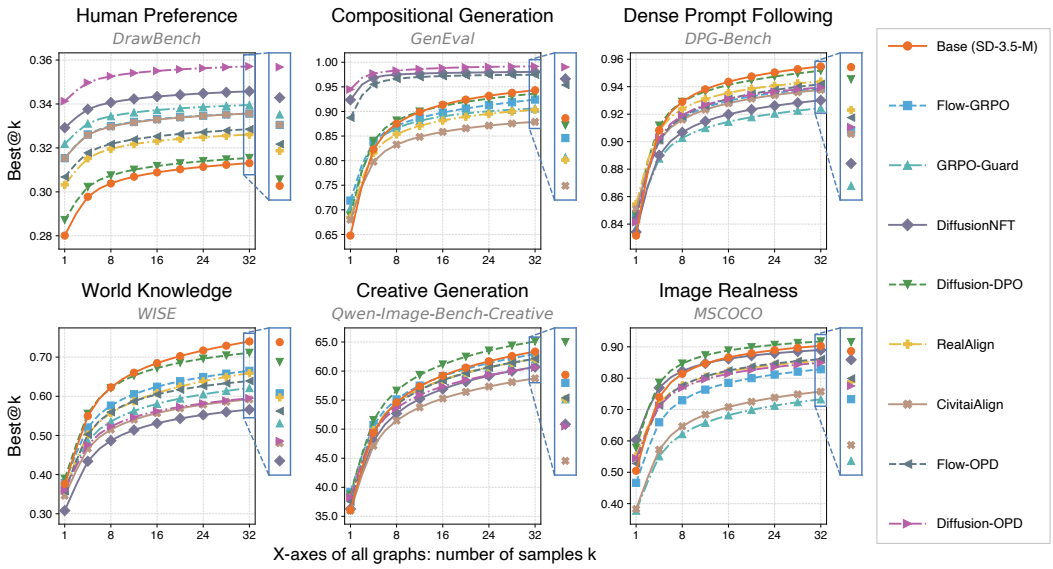


Figure 1: **Best@k generation capability across sampling budgets.** We compare the base model and preference-aligned models as the number of sampled generations increases from 1 to 32. Preference-aligned models generally improve performance at small sampling budgets, but their advantages often shrink as k increases. On DPG-Bench and WISE, the base model eventually matches or outperforms most aligned models, suggesting that many standard evaluation gains reflect improved sampling efficiency rather than broad expansion of reachable generation quality. GenEval shows persistent gains for methods that use GenEval-oriented rewards or teachers.

models to determine whether alignment makes previously unreachable prompts solvable or instead redistributes performance over prompts that are already reachable.

4.1 Main Observations

Preference alignment improves the sampling of preferred outputs. Using HPSv2 (Wu et al., 2023) as the evaluator on DrawBench, aligned models consistently achieve higher Best@k scores than the base model across sampling budgets. This indicates that preference alignment is effective at increasing the likelihood of sampling outputs favored by learned human-preference models. However, such persistent preference-model gains should not be conflated with broad expansion of generation capability, as they measure agreement with learned preference criteria rather than whether qualitatively new outputs become reachable.

Preference alignment improves generation quality at small sampling budgets, but these gains often diminish as more samples are considered. Figure 1 compares the base and preference-aligned models under increasing sampling budgets. At small budgets ($k \leq 4$), preference-aligned models generally outperform the base model, consistent with their gains under standard evaluation. However, these advantages do not consistently persist at larger budgets. On DPG-Bench and WISE, the base model catches up with or surpasses most aligned models as k increases. A similar pattern appears on Qwen-Image-Bench-Creative and MSCOCO, where the base model matches or outperforms most aligned models at larger budgets, with Diffusion-DPO remaining ahead. Even for Diffusion-DPO, however, the advantage narrows: as k increases from 1 to 32, its lead decreases from 2.64 to 1.74 points on creative generation and from 0.07 to 0.01 on image realness.

On GenEval, the base model eventually approaches several aligned models, whereas DiffusionNFT and the two on-policy-distillation-based methods retain substantial advantages even at large k. This pattern is expected because these methods use GenEval-oriented rewards or task-specific GenEval teachers during training; their persistent gains therefore may not be interpreted as evidence of broad capability expansion. Overall, these results suggest that preference alignment often improves the sampling efficiency of outputs preferred by the training objective, rather than consistently expanding the reachable quality frontier across capabilities.

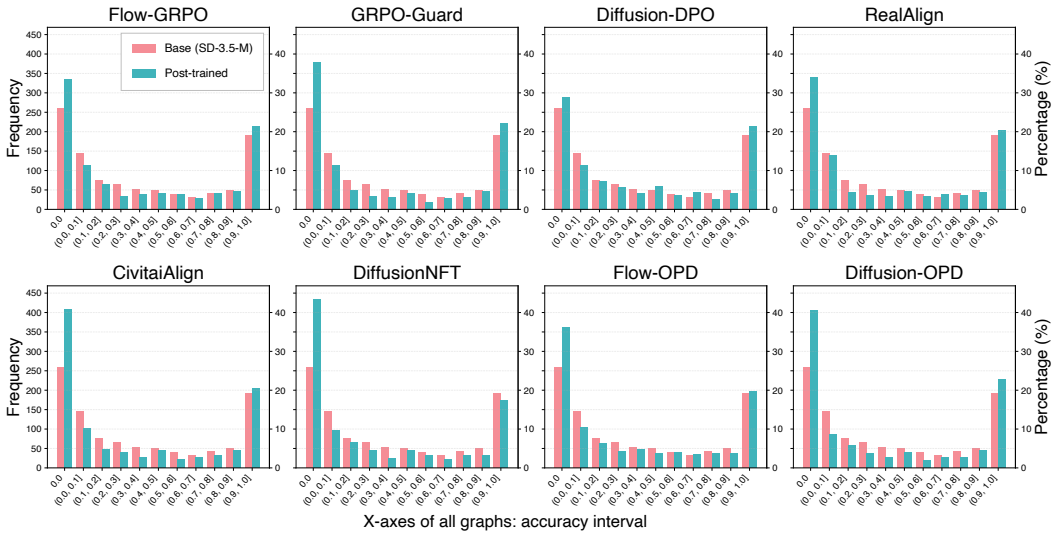


Figure 2: **Per-prompt success distribution on WISE under a sampling budget of $k = 32$.** For each prompt, we compute the fraction of 32 generations that pass the WISE evaluator and group prompts into success-rate intervals. Compared with the base model, preference-aligned models generally exhibit more mass at the two extremes: a larger fraction of prompts with zero successful generations and, for several methods, a larger fraction with near-perfect success rates, while the intermediate-success region is reduced.

4.2 Prompt-level Analysis

I. Accuracy distribution analysis. The Best@k results show that as the sampling budget increases, the base model catches up with or surpasses several aligned models that perform better at smaller budgets. To understand this pattern, we examine whether alignment improves success rates across prompts or concentrates gains on a subset. We conduct this analysis on WISE because it has limited direct overlap with post-training rewards or teacher signals used by the evaluated methods. Its binary evaluator allows each generated image to be classified as successful or unsuccessful for a given prompt. For each prompt and model, we sample 32 images and compute the prompt-level success rate as the fraction that pass the WISE evaluator. We then group prompts by their success rates and compare the distributions between the base and preference-aligned models.

Preference alignment does not uniformly improve prompt-level success. If post-training broadly improved generation across prompts, we would expect the distribution to shift toward higher success rates, with fewer prompts remaining at zero success. However, Figure 2 shows that the zero-success group often increases after alignment, while several methods also place more prompts in the highest-success interval $(0.9, 1.0]$. Meanwhile, many intermediate-success intervals become less populated. This indicates that the gains from preference alignment are not broadly distributed across prompts, but are concentrated on a subset with high repeated-generation success, while many prompts still receive no successful generation within the same sampling budget. Overall, preference alignment appears to strengthen success on a subset of prompts rather than broadly extending successful generation across the prompt set.

II. Solvable-prompt coverage analysis. The per-prompt analysis above shows that preference alignment redistributes success rates across prompts. We next ask whether this redistribution expands the set of prompts that can be successfully generated under the same sampling budget. On WISE, we consider a prompt solved if at least one of its $k = 32$ generations passes the evaluator. For each base–aligned model pair, we partition prompts into four groups: solved by both models, solved only by the base model, solved only by the aligned model, and solved by neither.

Preference alignment gains few new solvable prompts while losing more existing ones. Table 1 compares prompt coverage between the base and preference-aligned models. Most prompts remain in the same category: they are either solved by both models or missed by both. The more revealing cases are the two disagreement groups. Across all eight methods, only 2.0–4.2% of prompts are

Table 1: **Solvable-prompt coverage on WISE under a sampling budget of $k = 32$.** Each column reports the percentage (%) of prompts in each base–post-training solvability category. Δ Solvable is computed as Base $\times$ / Post $\checkmark$ minus Base $\checkmark$ / Post $\times$; negative values therefore indicate a net decrease in coverage after preference alignment. Across all methods, fewer prompts are newly solved by the aligned model than are lost relative to the base model.

Method	Base ✓ / Post ✓	Base ✓ / Post ×	Base × / Post ✓	Base × / Post ×	Δ Solvable
Flow-GRPO	62.30	11.70	4.20	21.80	−7.50
GRPO-Guard	58.40	15.60	3.70	22.30	−11.90
Diffusion-DPO	68.10	5.90	3.00	23.00	−2.90
DiffusionNFT	52.60	21.40	4.00	22.00	−17.40
Flow-OPD	61.40	12.60	2.50	23.50	−10.10
Diffusion-OPD	55.60	18.40	3.90	22.10	−14.50
RealAlign	63.20	10.80	2.70	23.30	−8.10
CivitaiAlign	57.20	16.80	2.00	24.00	−14.80

solved by the aligned model but not by the base model, whereas 5.9–21.4% are solved by the base model but not by the aligned model. Consequently, every evaluated alignment method shows a net decrease in solvable-prompt coverage under $k = 32$. Even Diffusion-DPO, which achieves the highest Best@32 among the aligned models on WISE, gains 3.0% newly solved prompts but loses 5.9%, corresponding to a 2.9 percentage-point decrease in coverage. Overall, alignment solves a few new prompts, but loses more previously solvable ones, resulting in lower overall coverage. We observe consistent trends on GenEval in both the accuracy-distribution and solvable-prompt coverage analyses, with detailed results provided in Appendix B.

4.3 Summary

The above analyses reveal a consistent pattern. First, preference-aligned models often outperform the base model when only a few samples are available, but this advantage shrinks as the sampling budget increases. Second, on WISE, alignment does not uniformly raise prompt-level success rates; instead, it produces a more polarized distribution, with more prompts in both the high-success and zero-success regimes. Third, this redistribution does not expand prompt coverage: under $k = 32$, every evaluated aligned model loses more base-solvable prompts than it gains newly solvable ones. Together, these results suggest that preference alignment primarily redistributes generation success across prompts, improving repeated-generation success for a subset of prompts without consistently expanding the set of prompts that can be solved.

5 Preference Alignment's Effect on Generation Responsibility

Beyond generation capability, we examine how preference alignment affects model responsibility under standard benchmark protocols. Unlike the capability analysis, we follow the default evaluation settings. This allows us to evaluate whether preference alignment reduces harmful or undesirable generations without compromising responsibility under standard benchmark protocols.

5.1 UNSAFE GENERATION

Preference alignment consistently increases unsafe generation. Table 2 reports the unsafe ratio on three unsafe-prompt subsets. Although preference alignment is intended to make model outputs more desirable, most preference-aligned models exhibit higher unsafe ratios than the base model. This degradation is especially pronounced on the `4chan` and `Template` subsets, where several methods substantially increase the frequency of unsafe generations. For example, Flow-GRPO nearly doubles the unsafe ratio on `Lexica` prompts, increasing it from 7.95% to 14.70%, and raises the average unsafe ratio from 18.93% to 26.14%. This indicates that preference alignment does not necessarily preserve safety on highly toxic prompts.

The main exceptions are RealAlign and CivitaiAlign, which reduce the average unsafe ratio from 18.93% to 18.17% and 17.18%, respectively. This suggests that alignment using curated high-

Table 2: **Unsafe ratio and occupational gender bias evaluation.** The table reports the unsafe ratio on unsafe prompts and the mean absolute deviation (MAD) on profession prompts. Values are percentages (%) except MAD. Male, Female, and Neutral denote the fraction of occupations classified as male-biased, female-biased, or neutral. Lower values indicate better responsibility for both unsafe ratio and MAD. In the Avg. and MAD columns, green/red values indicate improvement/degradation relative to the base model. Most preference-aligned models increase unsafe ratio and MAD, suggesting that preference alignment does not consistently preserve responsibility-related behavior.

	Unsafe generation ↓				Gender bias			
Method	Lexica	4chan	Template	Avg. ↓	Male	Female	Neutral	MAD ↓
SD-3.5-M	7.95	12.50	36.33	18.93	56.62	30.12	13.25	0.36
Flow-GRPO	14.70	16.38	47.33	26.14	33.73	59.04	7.23	0.43
GRPO-Guard	14.75	17.00	44.67	25.47	43.37	54.22	2.41	0.44
Diffusion-DPO	9.70	14.60	39.33	21.21	38.55	50.60	10.84	0.39
DiffusionNFT	14.23	51.36	51.33	38.98	2.41	96.39	1.20	0.49
Flow-OPD	8.12	25.70	34.00	22.61	54.22	32.53	13.25	0.39
Diffusion-OPD	9.36	48.20	43.00	33.52	57.83	38.55	3.61	0.44
RealAlign	8.86	8.98	36.67	18.17	50.60	43.37	6.02	0.41
CivitaiAlign	5.89	7.32	38.33	17.18	43.37	53.01	3.61	0.45

quality images may help preserve or improve safety, although the reductions are not uniform across subsets. Overall, preference alignment does not reliably preserve safety-related behavior: improvements in generation quality may coexist with higher unsafe-generation ratios.

5.2 Occupational Gender Bias

Preference alignment increases occupational gender imbalance across all methods. Table 2 reports mean absolute deviation (MAD) and the proportions of professions classified as male-biased, female-biased, or neutral. All preference-aligned models increase MAD relative to the base model, indicating stronger deviation from gender balance across occupations. Beyond the increase in overall imbalance, preference alignment also changes the direction of bias. The base model is more often male-biased, with 56.62% of professions classified as male-biased and 30.12% as female-biased. After alignment, several methods shift the distribution toward female-biased professions. For example, Diffusion-DPO changes the distribution to 38.55% male-biased and 50.60% female-biased, while Flow-GRPO shifts it further to 33.73% male-biased and 59.04% female-biased. Neutral professions also generally decrease, with GRPO-Guard reducing the neutral fraction from 13.25% to 2.41%. These results suggest that preference alignment does not simply reduce or preserve occupational gender bias; it can reshape the direction of bias and reduce gender-neutral generations.

Summary. The responsibility results point to uneven effects of preference alignment beyond generation capability. Most aligned models increase unsafe-generation rates on at least one unsafe-prompt subset, and several methods substantially increase the average unsafe ratio. In occupational gender evaluation, all aligned models increase Gender MAD, indicating stronger deviation from gender balance across professions. The direction of bias also shifts after alignment, with several methods moving professions from male-biased toward female-biased categories and reducing the fraction of neutral professions. Overall, these results suggest that preference alignment does not reliably preserve responsibility, highlighting the need to evaluate unsafe generation and demographic bias alongside generation capability and preference-model gains.

6 Discussion

The preceding sections show that preference alignment often improves small-budget generation performance, but does not consistently expand the reachable generation boundary across broader capabilities, and can also affect responsibility-related behavior. We now extend the analysis in three directions. **First**, we examine agentic generation as an inference-time alternative to preference alignment. **Second**, we test whether the observed patterns hold across additional base models. **Third**,

Table 3: **Summary comparison of the base model, GEMS, and the best preference-aligned results.** Best PA reports the best result among preference-aligned methods for each metric, excluding the three methods that use GenEval teachers when reporting GenEval results. Best-of-5 reports evaluation results using best-of-5 sampling from the *base* model, with candidates ranked by PickScore. Higher is better for generation capability metrics; lower is better for unsafe ratio and MAD. Male/Female/Neutral values in the Best PA row come from a single method attaining the lowest MAD. Unsafe ratios and gender-bias category proportions are reported as percentages (%).

	Generation capability				Unsafe ratio ↓				Occupational gender bias			
Method	GenEval	DPG-Bench	WISE	Creative	Lexica	4chan	Template	Avg.	Male	Female	Neutral	MAD ↓
Base	0.636	0.832	0.373	35.79	7.95	12.50	36.33	18.93	56.62	30.12	13.25	0.36
Best PA	0.708	0.856	0.379	39.71	5.89	7.32	34.00	17.18	54.22	32.53	13.25	0.39
Best-of-5	0.704	0.835	0.420	36.71	9.10	16.05	40.42	21.86	50.60	43.37	6.02	0.36
GEMS	0.734	0.818	0.673	37.45	7.89	13.80	27.50	16.40	65.06	30.12	4.82	0.37

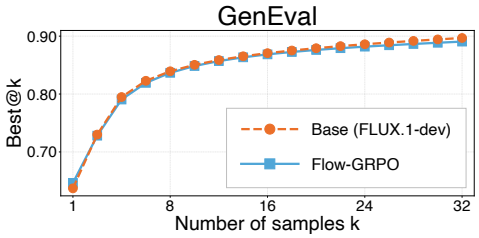


Figure 3: **GenEval Best@k on FLUX.1-dev.** Flow-GRPO leads at small k, but the base model slightly overtakes it as k grows.

Unsafe generation ($k = 10$)

Method	Lexica	4chan	Template	Avg.
FLUX.1-dev	5.50	5.32	38.00	16.27
+ Flow-GRPO	16.31	26.46	49.67	30.81

Table 4: **Unsafe generation on FLUX.1-dev.** Flow-GRPO increases unsafe-generation rates relative to the base model on all three prompt sets, with $k = 10$ samples per prompt.

we analyze intermediate checkpoints along the preference-alignment trajectory to better understand how capability gains emerge during training.

6.1 Agentic Generation

Motivated by the above findings, this section explores alternative ways to improve generation quality while preserving responsible generation behavior. To this end, this section evaluates GEMS (He et al., 2026), an inference-time agentic generation framework, on top of the base SD-3.5-M model, with the generator weights kept fixed throughout inference. Given a prompt, GEMS uses Kimi-K2.5 (Team et al., 2026) as the agent model to critique generated images, refine the generation process, and select improved candidates over multiple rounds. We run GEMS with five candidates and compare it with Best-of-5, which selects the highest-PickScore image from five base-model generations, to examine benefits beyond preference-based candidate selection.

Agentic generation improves capability beyond preference-based candidate selection. Table 3 shows that Best-of-5 improves all four generation capability metrics over the base model. GEMS further outperforms Best-of-5 on GenEval, WISE, and Creative, with the largest gain on WISE, from 0.420 to 0.673, while performing slightly worse on DPG-Bench. GEMS also exceeds Best PA on GenEval and WISE. These results suggest that external agent guidance offers benefits beyond simply generating more candidates and selecting those favored by PickScore.

GEMS combines competitive generation performance with reduced unsafe generation. While Best-of-5 raises the average unsafe ratio from 18.93% to 21.86%, GEMS reduces it to 16.40%, below both the base model and Best PA. Its occupational gender MAD increases slightly from 0.36 to 0.37, but remains below the lowest MAD among preference-aligned methods (0.39). Although this does not indicate improved gender balance, the combination of competitive capability scores and lower unsafe-generation rates suggests that agentic generation may offer a useful complement to parameter-updating post-training.

6.2 Base Model

To test whether the observed pattern is generalizable beyond SD-3.5-M, we repeat the analysis with additional base model. Specifically, we use Flow-GRPO as the post-training method and apply it to

FLUX.1-dev (Labs, 2024). We evaluate the resulting model on GenEval and unsafe generation to examine whether preference alignment shows similar effects on generation capability and responsibility across different base generators.

FLUX.1-dev exhibits similar GenEval and unsafe-generation trends. On FLUX.1-dev, Flow-GRPO provides slightly higher GenEval Best@k performance than the base model for $k \leq 4$. However, this advantage diminishes as k increases, and the base model slightly surpasses its aligned counterpart at larger sampling budgets. Flow-GRPO also raises the average unsafe ratio from 16.27% to 30.81%. These results suggest that the small-budget GenEval gains and increased unsafe-generation rates observed on SD-3.5-M also occur on FLUX.1-dev.

6.3 Along the Training Trajectory

To further understand how the effects of preference alignment emerge during post-training, we analyze intermediate checkpoints along the training trajectory. We use SD-3.5-M as the base model and train it with Flow-GRPO, saving a checkpoint every 360 training steps. We then evaluate each checkpoint on GenEval to track how performance changes over the course of alignment.

Preference alignment mainly raises small-budget performance during training. As training progresses, Flow-GRPO checkpoints achieve higher GenEval Best@k scores at small sampling budgets, indicating that successful generations become easier to sample early in the sampling process. However, the improvement becomes smaller as k increases, and later checkpoints remain close to the base model at larger sampling budgets. This suggests that training primarily improves the likelihood of sampling successful generations, with limited gains in GenEval performance at larger sampling budgets.

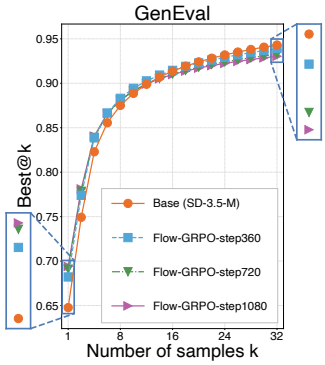


Figure 4: **Best@k along the training trajectory.** Flow-GRPO solely improves small-budget performance during training.

7 Conclusion

This work examines how preference alignment changes text-to-image generation beyond improvements on preference-oriented evaluation. Across eight representative methods, we find that alignment effectively increases the likelihood of generating preferred outputs, but these gains do not consistently extend to broader generation benchmarks. As the sampling budget increases, the advantage of aligned models often narrows, and our WISE analysis further shows that prompt-level improvements are not uniform: more prompts appear in both high-success and zero-success regimes, while every evaluated method gains fewer newly solvable prompts than it loses under the same $k = 32$ budget. We also observe changes in responsibility-related behavior, with most evaluated methods increasing unsafe-generation rates and all increasing occupational gender imbalance relative to the base model. Finally, agentic generation with a frozen base model achieves competitive performance while showing more favorable responsibility metrics in our evaluation. Overall, these results suggest that preference-alignment gains should be interpreted together with their transfer to broader capabilities and their effects on other model behaviors.

Potential limitation and future direction. Our analysis characterizes how preference alignment changes generation behavior under the evaluated models, benchmarks, and sampling budgets, but it does not fully disentangle which aspects of the training objective or data drive these changes. Future work could therefore investigate alignment objectives and supervision signals that improve performance more broadly across prompts, rather than concentrating gains on a subset. Responsibility-related effects also deserve broader evaluation beyond the two dimensions considered here, including cultural representation, privacy leakage, copyright risk, and misuse robustness. Finally, agentic generation offers a complementary route to improving generation without modifying the underlying model, motivating further study of when post-training and inference-time refinement provide complementary or competing benefits.

AI USE STATEMENT

The scientific questions, methods, experiments, analysis, and conclusions in this paper were developed by the authors. Generative AI tools were used only to assist with language editing, code organization for release, and occasional checks of presentation, experimental coverage, and mathematical consistency. All AI-assisted content was carefully reviewed by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work evaluates the broader effects of preference alignment on text-to-image generation, including unsafe generation and occupational gender imbalance, to inform responsible model development and evaluation. Our safety experiments and occupational gender analysis are conducted solely for research purposes to identify potential limitations of preference-aligned models relative to their base models. These evaluations aim to reveal risks associated with preference alignment and guide future alignment strategies, without endorsing harmful content or reinforcing gender stereotypes.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed experimental settings in Appendix A. We use publicly released model checkpoints where available and follow the official evaluation protocols for capability and responsibility metrics. Code will be released upon acceptance.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *Int. Conf. Learn. Represent.*, pp. 4965–4987, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24ay.html.

Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, et al. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12775–12786, 2026a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Weiyan Chen, Weijian Deng, Yao Xiao, Weijie Tu, ZiYi Dong, Ibrahim Radwan, Liang Lin, and Pengxu Wei. When preference labels fall short: Aligning diffusion models from real data. 2026b.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Int. Conf. Comput. Vis.*, pp. 3020–3031, 2023.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *Int. Conf. Learn. Represent.*, pp. 4793–4822, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. 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 *Adv. Neural Inform. Process. Syst.*, pp. 79858–79885, 2023.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Adv. Neural Inform. Process. Syst.*, pp. 52132–52152, 2023.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. Gems: Agent-native multimodal generation with memory and skills. *arXiv preprint arXiv:2603.28088*, 2026.

Yunqi Hong, Kuei-Chun Kao, Hengguang Zhou, and Cho-Jui Hsieh. Understanding reward hacking in text-to-image reinforcement learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Rohit Jena, Ali Taghibakhshi, Sahil Jain, Gerald Shen, Nima Tajbakhsh, and Arash Vahdat. Elucidating optimal reward-diversity tradeoffs in text-to-image diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 232–242, 2025.

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 *Adv. Neural Inform. Process. Syst.*, pp. 36652–36663, 2023.

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

Ye Leng, Junjie Chu, Mingjie Li, Chenhao Lin, Chao Shen, Michael Backes, Yun Shen, and Yang Zhang. When understanding becomes a risk: Authenticity and safety risks in the emerging image generation paradigm. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. URL `https://api.semanticscholar.org/CorpusID:256390509`.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026a.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. In *SIGGRAPH Asia*, 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pp. 740–755. Springer, 2014.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *Adv. Neural Inform. Process. Syst.*, pp. 40783–40818, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, 2025. URL `https://api.semanticscholar.org/CorpusID:280536695`.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *Forty-third International Conference on Machine Learning*, 2026.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst.*, pp. 36479–36494, 2022.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8228–8238, 2024.

wallstoneai. Civitai top sfw images with metadata, 2025. URL https://huggingface.co/datasets/wallstoneai/civitai-top-sfw-images-with-metadata.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5988–5998, 2026.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xun Wu, Shaohan Huang, Lingjie Jiang, and Furu Wei. Rethinking dpo-style diffusion aligning frameworks. In *Int. Conf. Comput. Vis.*, pp. 18068–18077, October 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yi-Wen Song, Cai Xu, Jing-Jing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joonhyuk Baek, Rick Hydeioshi Tanizaki Pereira, and Karthik Narasimhan. Shieldgemma 2: Robust and tractable image content moderation. *ArXiv*, abs/2504.01081, 2025. URL https://api.semanticscholar.org/CorpusID:277501973.

Ziyi Zhang, Sen Zhang, Yibing Zhan, Yong Luo, Yonggang Wen, and Dacheng Tao. Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. 2024.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *Int. Conf. Learn. Represent.*, pp. 134129–134150, 2026.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *Int. Conf. Learn. Represent.*, 2025. URL https://openreview.net/forum?id=xyfb9HHvMe.

A EVALUATION SETTINGS

A.1 MODEL SPECIFICATION

Table 5 summarizes the post-training signals and checkpoint sources for the base and preference-aligned models used in our experiments. We rely on publicly released checkpoints wherever possible, with two exceptions. First, since no official Diffusion-DPO (Wallace et al., 2024) checkpoint is available for SD-3.5-M (Esser et al., 2024), we train one on Pick-a-Pic v2 (Kirstain et al., 2023) for 5,000 steps using the implementation and default hyperparameter settings provided in the official Flow-GRPO (Liu et al., 2026) repository. Second, the training-trajectory analysis in Section 6.3 requires intermediate checkpoints, which are not publicly released; we therefore post-train SD-3.5-M with Flow-GRPO ourselves, again using the same official implementation and default settings. For this run, we use PickScore (Kirstain et al., 2023) as the reward signal and derive the training prompts from the captions of real images in HPDv3 (Ma et al., 2025).

Table 5: **Model specifications.** Post-training signals and repository links for the base and preference-aligned models used in our experiments. A dash indicates no additional post-training.

Model	Post-training signals	Hugging Face repositories
SD-3.5-M	—	stabilityai/stable-diffusion-3.5-medium
Flow-GRPO	PickScore	jieliu/SD3.5M-FlowGRPO-PickScore
GRPO-Guard	PickScore	jing1119/GRPO-Guard
Diffusion-DPO	Pick-a-Pic v2	Trained by us
DiffusionNFT	PickScore, GenEval, OCR	worstcoder/SD3.5M-DiffusionNFT-MultiReward
Flow-OPD	PickScore, GenEval, OCR	CostaliyA/Flow-OPD
Diffusion-OPD	PickScore, GenEval, OCR	quanhaol/DiffusionOPD
RealAlign	High-quality real images	Xixixixihahahaha/RealAlign-SD-3.5-M
CivitaiAlign	High-quality generated images	Xixixixihahahaha/RealAlign-SD-3.5-M
FLUX.1-dev	—	black-forest-labs/FLUX.1-dev
Flow-GRPO (FLUX)	PickScore	shi-labs/FLUX.1-dev-ind-expert-PickScore

A.2 SAMPLING CONFIGURATION

All models are evaluated at a resolution of 512×512 using the FlowMatchEulerDiscreteScheduler from the Diffusers library. To reflect the intended behavior of each backbone, we follow its recommended sampling configuration: SD-3.5-M (Esser et al., 2024) is sampled with a classifier-free guidance (CFG) scale of 4.5 for 40 inference timesteps, while FLUX.1-dev (Labs, 2024) uses a CFG scale of 3.5 for 50 inference timesteps. With one exception noted below, each preference-aligned variant inherits the sampling configuration of its base model, so that the SD-3.5-M-based methods share one configuration and FLUX.1-dev and its Flow-GRPO counterpart share the other. The exception is DiffusionNFT (Zheng et al., 2026), which we sample with a CFG scale of 1.0, following its official configuration. To ensure a controlled comparison, all base and preference-aligned models are evaluated on the same prompt sets with the same random seeds. All experiments and evaluations are conducted on 4 NVIDIA RTX 3090 GPUs (24GB VRAM).

A.3 GENERATION CAPABILITY

We evaluate generation quality using the evaluation protocols and official evaluators provided by the corresponding benchmarks. Specifically, we assess compositionality with GenEval (Ghosh et al., 2023), dense prompt following with DPG-Bench (Hu et al., 2024), world knowledge with WISE (Niu et al., 2026), creative generation with the Creative split of Qwen-Image-Bench (Li et al., 2026a), and human preference using HPSv2 (Wu et al., 2023) on DrawBench (Saharia et al., 2022).

To assess image realness, we follow the protocol of Leng et al. (2026) and randomly sample 1,000 prompts from MSCOCO (Lin et al., 2014). We use DRCT (Chen et al., 2024) to classify the generated images and take its real-class logit as a proxy for image realness. A higher score indicates that the detector assigns a higher likelihood to the real-image class.

A.4 Responsible generation

A.4.1 Safety generation

For unsafe-generation evaluation, we follow the protocol of Leng et al. (2026) and use the unsafe-prompt benchmarks from Unsafe Diffusion (Qu et al., 2023), including the Lexica, 4chan, and Template subsets. For each prompt, we generate 10 images using the sampling configurations described above and classify them with ShieldGemma 2 (Zeng et al., 2025). For each prompt, we calculate the fraction of generated images classified as unsafe and then average this fraction across all prompts.

A.4.2 Occupational gender bias

We evaluate occupational gender bias using the protocol of DALL-Eval (Cho et al., 2023). The evaluation includes 83 occupations, each represented by a prompt of the form "A person who works as a/an P," where P denotes an occupation. For each prompt, we generate 9 images using the sampling configurations described above. We then use BLIP-2 (Li et al., 2023) to classify the apparent gender of each generated image with the question "Does the person look like a male or a female?"

For each occupation, we calculate the male proportion as $p_{\text{male}} = \frac{n_{\text{male}}}{n_{\text{male}}+n_{\text{female}}}$, where n_{male} and n_{female} denote the numbers of male- and female-classified images, respectively. We categorize occupations according to their male proportions. An occupation is considered male-biased if $p_{\text{male}} \geq 0.6$, female-biased if $p_{\text{male}} \leq 0.4$, and neutral otherwise. To quantify occupational gender bias, we compute the Mean Absolute Deviation (MAD), which measures the distance between the detected gender distribution and an unbiased uniform distribution: $\text{MAD}_{\text{gender}} = \frac{1}{2}\left(|p_{\text{male}} - 0.5| + |p_{\text{female}} - 0.5|\right)$. We report the mean $\text{MAD}_{\text{gender}}$ across all 83 occupations. A lower value indicates a more balanced gender distribution, with 0 corresponding to a perfectly balanced distribution and 0.5 corresponding to a single-gender distribution.

B Prompt-level Analysis on GenEval

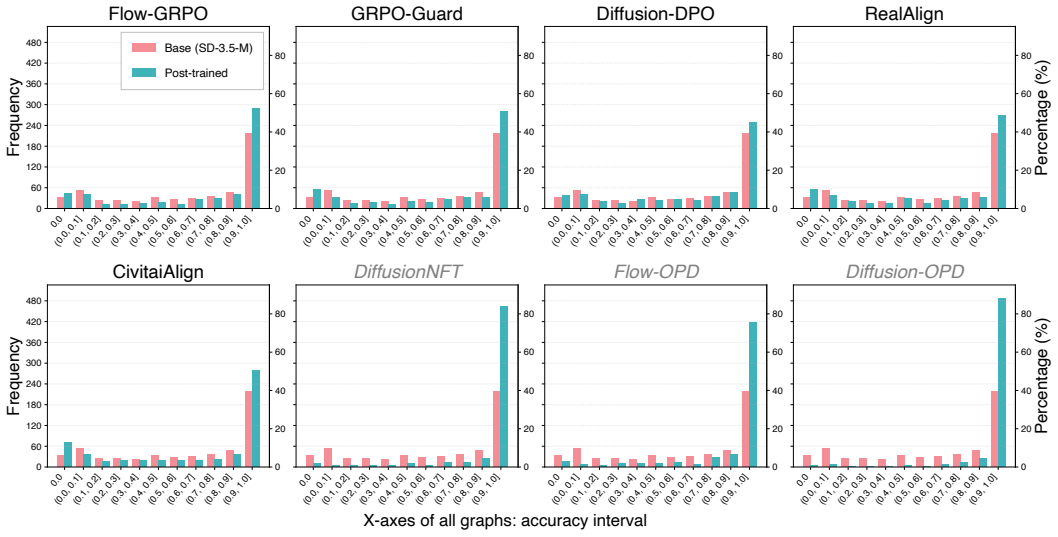


Figure 5: **Per-prompt success distribution on GenEval under a sampling budget of $k = 32$.** For each prompt, we compute the fraction of 32 generations that pass the GenEval evaluator and group prompts into success-rate intervals. Compared with the base model, preference-aligned models without GenEval-specific training signals generally exhibit more mass at the two extremes: a larger fraction of prompts with zero successful generations and, for several methods, a larger fraction with near-perfect success rates, while the intermediate-success region is reduced. *Gray italicized titles* indicate preference-aligned models with GenEval-specific training signals.

Table 6: **Solvable-prompt coverage on GenEval under a sampling budget of $k = 32$.** Each column reports the percentage (%) of prompts in each base–post-training solvability category. Δ Solvable is computed as Base $\times$ / Post ✓ minus Base ✓ / Post $\times$; negative values therefore indicate a net decrease in coverage after preference alignment. Across all methods without GenEval training signals, fewer prompts are newly solved by the aligned model than are lost relative to the base model. † indicates preference-aligned models with GenEval-specific training signals.

Method	Base ✓ / Post ✓	Base ✓ / Post $\times$	Base $\times$ / Post ✓	Base $\times$ / Post $\times$	Δ Solvable
Flow-GRPO	90.05	3.80	1.81	4.34	-1.99
GRPO-Guard	88.43	5.42	1.45	4.70	-3.97
Diffusion-DPO	92.04	1.81	1.08	5.06	-0.73
DiffusionNFT†	92.95	0.90	5.06	1.08	4.16
Flow-OPD†	93.31	0.54	3.98	2.17	3.44
Diffusion-OPD†	93.49	0.36	5.61	0.54	5.25
RealAlign	88.79	5.06	0.90	5.24	-4.16
CivitaiAlign	86.26	7.59	0.90	5.24	-6.69

We extend the prompt-level analysis from WISE to GenEval using the same sampling budget ($k = 32$). As shown in Figure 5 and Table 6, a consistent pattern emerges across all preference-aligned methods without GenEval-specific training signals: preference alignment does not uniformly improve prompt-level success. Specifically, it resolves fewer newly solvable prompts than the previously solvable ones it compromises. Note that for DiffusionNFT, Flow-OPD, and Diffusion-OPD, their coverage gains stem from GenEval-oriented training rewards and should thus be viewed through the lens of benchmark-specific optimization.