HIGH-QUALITY DEMONSTRATIONS IN DIFFUSION REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) improves diffusion models by rewarding generations that satisfy task objectives. Yet task rewards capture only part of visual quality, while the model's generations may continue to fall short of task requirements. High-quality external prompt–image pairs, which we call demonstrations, provide holistic visual supervision beyond what task rewards capture. To incorporate this supervision into RL, we introduce Demonstration-Augmented NFT (Demo-NFT), which combines two sources of image targets in DiffusionNFT's objective: prompt-matched external demonstrations and native samples generated by the diffusion model being optimized. Their rewards jointly determine the advantages used to weight these targets. However, large reward gaps can skew this weighting toward demonstrations. Native-referenced reward capping calibrates demonstration rewards against those of native samples. With 5% demonstration targets, Demo-NFT improves QJudge, an auxiliary metric of image quality, by 8.84 points on average over NFT across FLUX and SD3.5 on compositional generation and text rendering, while achieving better or comparable task performance. Using only the task reward, Demo-NFT also outperforms a multi-reward NFT baseline combining task and HPSv3 rewards on the task metric, QJudge, and PickScore in all four settings. We further examine supervised mixing and shared SFT initialization, and extend the evaluation to Qwen-Image and SenseNova. These results show that Demo-NFT improves multiple aspects of image quality using a single task reward.

1 INTRODUCTION

Reinforcement learning (RL) adapts diffusion models (Ho et al., 2020; Rombach et al., 2022) by rewarding generations that satisfy task objectives, such as compositional accuracy and text rendering (Black et al., 2023; Fan et al., 2023; Liu et al., 2025; Zheng et al., 2025). The learning signal, however, depends on both what the task reward measures and what the model generates. Task rewards measure selected attributes rather than every aspect of a desirable image (Hu et al., 2023). An OCR reward, for example, evaluates text accuracy but offers limited guidance on typography, visual hierarchy, and composition. Optimizing the measured attributes can therefore improve task scores without preserving other aspects of image quality (Black et al., 2023; Lee et al., 2023; Ye et al., 2025). Adding reward functions broadens coverage, but some aspects of image quality remain unmeasured. The examples available for learning also matter: the model's generations may continue to fall short of task requirements, exhibiting counting or spatial-relation errors (Ghosh et al., 2023; Huang et al., 2023; Kamath et al., 2025) and long-form text-rendering errors (Geng et al., 2025).

A well-designed magazine cover, for example, combines accurate text with typography, visual hierarchy, and coherent layout. We call high-quality external prompt–image pairs that illustrate such desired outputs demonstrations. They provide holistic visual supervision: complete image targets convey task execution together with typography, layout, and other aspects of visual quality that the task reward does not fully capture. Rather than specifying an additional score for each property, demonstrations show how these properties can be realized together. When errors persist in the model's generations, they can also provide more accurate examples of task execution.

To incorporate this supervision into RL, we introduce Demonstration-Augmented NFT (Demo-NFT). For each prompt, we combine native samples generated by the diffusion model being op-

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.

timized with prompt-matched external demonstrations. We build on DiffusionNFT (Zheng et al., 2025), whose forward-process objective learns from clean-image targets without requiring their generation trajectories. This interface allows prompt-matched demonstrations to replace a fraction of native training targets. Both native samples and demonstrations receive the task reward. Their advantages are estimated jointly and determine how each image target contributes to the shared NFT objective. Demonstrations can be curated real or synthetic images. Specifically, our experiments use GPT-image-2 outputs.

Joint weighting introduces a calibration issue when demonstrations score well above native samples. We therefore introduce native-referenced reward capping, limiting each demonstration reward to the highest retained native-sample reward for the same prompt before advantage estimation. This calibrates the two sources' relative influence without changing their image targets.

We evaluate compositional generation and text rendering on FLUX 4B (Black Forest Labs, 2026) and SD3.5 (Esser et al., 2024), with additional experiments on Qwen-Image (Wu et al., 2025) and SenseNova (Diao et al., 2026).

Our contributions are twofold. (1) We introduce Demonstration-Augmented NFT (Demo-NFT), which integrates native samples and prompt-matched external demonstrations into a shared DiffusionNFT objective. Joint advantage estimation and native-referenced reward capping coordinate learning from both sources, while our analysis distinguishes image-target supervision from reward weighting. (2) Across four main model–task settings, using demonstrations for only 5% of training targets improves all four auxiliary metrics over NFT, with better or comparable task metrics. Using a single task reward, Demo-NFT also outperforms the task-plus-HPSv3 multi-reward NFT baseline on the task metric, QJudge, and PickScore in all four settings. Auxiliary-metric gains persist after shared SFT initialization, supporting continued demonstration use during RL. Against supervised mixing, Demo-NFT achieves higher task scores in three of four settings with a lower demonstration ratio.

2 Related Work

We relate our approach to diffusion post-training (Prabhudesai et al., 2023; Wallace et al., 2024) and learning from external examples.

Diffusion post-training. DDPO (Black et al., 2023) and DPOK (Fan et al., 2023) optimize denoising policies through policy gradients, with DPOK adding KL regularization, while DRaFT (Clark et al., 2023) differentiates rewards through sampling. Flow-GRPO (Liu et al., 2025) adapts group-relative policy optimization to flow models through stochastic sampling and studies KL regularization for preserving quality during reward optimization. DiffusionNFT (Zheng et al., 2025) instead learns from clean-image targets through a forward-process objective with positive and negative reinforcement signals. We retain this objective and extend its target distribution with external demonstrations.

Learning from external supervision. High-quality external prompt–image pairs provide image-level demonstrations of desired outputs. DDRL (Ye et al., 2025) combines reward maximization with data-anchored diffusion training, motivated by forward-KL regularization. Its video generation and image text-rendering experiments show that even task-imperfect synthetic images provide useful visual regularization. In language-model reasoning, LUFFY (Yan et al., 2025) combines external reasoning demonstrations with on-policy rollouts. For image generation, the concurrent preprint PortraitGen (Li et al., 2026b) incorporates real-image exemplars into GRPO groups using image inversion to recover sampling trajectories. Demo-NFT directly incorporates external images as clean targets in a shared reward-weighted NFT objective, without recovering their generation trajectories. The distinction from supervised fitting is whether demonstration rewards participate in the learning weights: Demo-NFT includes them in joint advantage estimation within the shared NFT objective.

3 Demonstration-Augmented NFT with Native-Referenced Reward Capping

As discussed in the Introduction, demonstrations are high-quality external prompt–image pairs that provide holistic visual supervision. Native samples are generated by the diffusion model being opti-

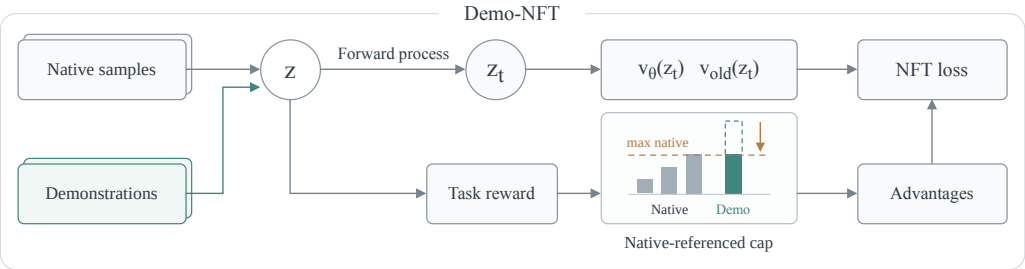


Figure 1: **Demo-NFT overview.** Native samples and high-quality external images paired with prompts (demonstrations) share NFT's forward-process objective. Native-referenced capping calibrates demonstration rewards before advantage estimation without changing image targets. Demo-NFT (w/o cap) omits this step. The forward-process layout follows DiffusionNFT (Zheng et al., 2025).

mized. Demo-NFT integrates both target sources through NFT and calibrates their relative weighting with native-referenced reward capping.

3.1 Constructing the training targets

Let p be a prompt, $r(x, p)$ its task reward, and $\mathcal{D}$ the set of demonstration pairs (p, x), with $\mathcal{D}_p$ containing the images paired with p. At each iteration, the sampling policy generates a group of images per prompt. We replace a fraction ρ of targets across the training batch with demonstrations drawn from $\mathcal{D}_p$ (Figure 1). The final group is $\mathcal{G}_p = \mathcal{N}_p \cup \mathcal{D}_p^{\text{used}}$, comprising retained native targets $\mathcal{N}_p$ and selected demonstrations $\mathcal{D}_p^{\text{used}}$.

The full demonstration set remains available for sampling. Both sources serve as clean-image targets under the same NFT objective.

3.2 Native-referenced reward capping

Under joint advantage estimation, demonstration rewards affect the weighting of both sources through their shared prompt baseline. Each prompt group retains native samples, whose highest reward defines the cap for its demonstrations. We apply this cap before computing advantages:

$$M_p = \max_{x \in \mathcal{N}_p} r(x, p), \qquad \widetilde{r}(x, p) = \begin{cases} \min\{r(x, p), M_p\}, & x \in \mathcal{D}_p^{\text{used}}, \\ r(x, p), & x \in \mathcal{N}_p. \end{cases} \tag{1}$$

Diagnostics and evaluation use original task rewards; advantage estimation uses capped rewards.

Let $\mathcal{B}$ denote the distributed training batch and $\sigma_{\mathcal{B}}$ the standard deviation of the rewards $\widetilde{r}$ used for advantage estimation. The evaluated configuration normalizes each reward using a prompt-specific mean and the shared batch standard deviation:

$$A_i = \frac{\widetilde{r}_i - \mu_{p_i}}{\max(\sigma_{\mathcal{B}}, \sigma_{\min})}, \qquad \mu_p = \frac{1}{|\mathcal{G}_p|} \sum_{x \in \mathcal{G}_p} \widetilde{r}(x, p), \qquad \sigma_{\min} = 0.1. \tag{2}$$

Advantages are clipped to $[-a, a]$, with $a > 0$, before computing the NFT branch weights. The ablation variant Demo-NFT (w/o cap) uses the same normalization with $\widetilde{r} = r$, providing a direct control for the cap.

3.3 Optimizing the mixed targets

We express NFT's positive and negative branches as clean-state predictions (Zheng et al., 2025). For a clean target z, we draw noise $\epsilon \sim \mathcal{N}(0, I)$ and form $z_t = (1 - t)z + t\epsilon$ (Lipman et al., 2022; Liu et al., 2022). Let v_θ and v_{old} denote the trainable and detached moving-reference velocity predictions at (z_t, t, p). Their corresponding clean-state predictions are $\widehat{z}_\theta = z_t - tv_\theta$ and $\widehat{z}_{\text{old}} = z_t - tv_{\text{old}}$.

With coefficient $\beta > 0$, the two branches are

$$\widehat{z}_\theta^{+} = (1 - \beta)\widehat{z}_{\text{old}} + \beta\widehat{z}_\theta, \qquad \widehat{z}_\theta^{-} = (1 + \beta)\widehat{z}_{\text{old}} - \beta\widehat{z}_\theta. \tag{3}$$

Both branches reconstruct the same clean target. Appendix A gives the equivalence to NFT's velocity-space formulation. The advantage determines their relative weights through $q_i = \frac{1}{2} + \mathrm{clip}(A_i, -a, a)/(2a)$. The NFT objective is

$$\mathcal{L}_{\mathrm{NFT}} = \frac{a}{\beta}\mathbb{E}_i \left[q_i \ell(\widehat{z}^{+}_{\theta,i}, z_i) + (1 - q_i)\ell(\widehat{z}^{-}_{\theta,i}, z_i) \right]. \tag{4}$$

For the main latent-space backbones, z is an encoded image latent. The expectation includes the sampled loss timesteps and noise. Here $\ell(u, z)$ is the mean squared reconstruction error divided by the detached mean absolute reconstruction error, with the denominator floored at 10^{-5}; both means are over latent coordinates and are computed separately for each target and branch. We also penalize squared velocity-prediction deviation from the frozen base model. Native and demonstration targets use Equation (4) with their respective advantages, so reward calibration directly changes their weighting within the NFT objective. Appendix A specifies the optimization coefficients.

Within each training batch, native samples and demonstrations for the same prompt share loss timesteps and added noise, aligning their forward-process perturbations.

Beyond scalar rewards. In Equation (4), the advantage determines how a target contributes to the NFT loss, while the target image determines the reconstruction signal. For a fixed batch of native samples, adding rewards changes their weighting; adding demonstrations also changes the images available for learning. For example, an OCR reward measures text accuracy, whereas a high-quality demonstration also supplies typography, layout, and visual hierarchy through its reconstruction target. Through joint advantage estimation, Demo-NFT integrates holistic visual supervision from demonstration images with task-reward feedback in a shared NFT objective. Appendix F formalizes this distinction through reward-weighted target distributions and the NFT gradient.

3.4 Effect on the learning signal

The shared positive affine normalization in Equation (2) preserves reward ordering, so $\widetilde{r}_d \le M_p$ gives the advantage ceiling

$$A^{\mathrm{cap}}_d \le \max_{x \in \mathcal{N}_p} A^{\mathrm{cap}}_x. \tag{5}$$

Above the highest retained native reward, demonstrations share its capped advantage; below it, rewards are unchanged. Capping thus calibrates their weights against retained native-sample rewards while preserving image content.

On the same mixed batch, capping lowers prompt baselines without changing native-sample rewards, preserving all nonnegative native-sample advantages (Appendix E). As the batch standard deviation can change, the numerical values of all native-sample advantages need not increase.

Figure 4(b) shows the paired advantage distributions, and Figure 4(c) their gradient consequences. Appendix E derives the paired gradient identity at matched parameters, targets, noise, and timesteps; gradient magnitudes depend on the branch gradients and their alignment.

4 Experimental Setup

To assess the roles of demonstration supervision and reward calibration, we compare native-only RL, demonstration-augmented RL, and alternative training strategies on compositional generation and text rendering.

4.1 Tasks and demonstration data

In our experiments, demonstration sets are prepared before training, kept unchanged during training, and shared across methods within each task.

Compositional generation. Training uses about 2,000 GenEval2-style prompts and about 8,000 GPT-image-2 demonstrations. We evaluate on a 128-prompt GenEval2 set, normalized to the training template without changing prompt meaning. Appendix C describes data preparation and overlap checks.

Long-form text rendering. We introduce LongText-Extend by expanding English examples from LongText-Bench (Geng et al., 2025), with 2,000 training prompts and 100 held-out evaluation

prompts. These specify multiple text fields, layout, and typography. Four GPT-image-2 demonstrations per training prompt yield 8,000 images. Tables and figures refer to this task as OCR.

4.2 Models and controls

FLUX.2-klein-base-4B (Black Forest Labs, 2026) and Stable Diffusion 3.5 Medium (Esser et al., 2024) provide the main comparisons on both tasks at 1024 resolution. NFT uses only native targets and the task reward. Parentheses in NFT table labels identify the reward configuration: OCR or GenEval2 alone, or the task reward together with HPSv3. Demo-NFT includes native-referenced reward capping and uses 5% demonstrations by default; its ablations remove capping or increase the ratio to 10%.

We implement DDRL's data-regularization strategy using NFT, denoted DDRL*: supervised flow matching applies to 10% demonstration targets and NFT to native targets. As a representative multi-reward baseline, we combine the task reward with the HPSv3 preference reward (Ma et al., 2025) under NFT. Each reward is normalized separately to obtain advantages, which are then averaged with equal weights. All NFT configurations use advantage clipping with $a = 5$. Appendices A and D specify the loss weighting and optimization details.

Qwen-Image-2512 (Wu et al., 2025) and SenseNova-U1.5-8B-MoT (Diao et al., 2026) extend the evaluation to additional backbones on both tasks. NFT and Demo-NFT use matched training setups within each backbone and task; Appendix A provides backend-specific settings.

4.3 Evaluation and checkpoint selection

All reported images use 1024 resolution, 50 sampling steps, and guidance scale 3.5 (Ho & Salimans, 2022) with the corresponding backend sampler. Within each task, methods share prompts and generation seeds.

For the main FLUX/SD3.5 comparisons, we select the checkpoint with the highest task score: Soft-TIFA AM for GenEval2 and OCR score for OCR. These are best observed test results, not validation-selected estimates. All auxiliary metrics use the same checkpoint as the task score. Appendices B and D record evaluation checkpoints, extension selections, and restoration semantics.

4.4 Task metrics and auxiliary metrics

Task metrics. GenEval2 uses Soft-TIFA arithmetic mean (AM), averaging question-level probabilities within each image and then across images (Kamath et al., 2025). For LongText-Extend, the OCR score combines character-similarity-weighted word recall and reading order.

Auxiliary metrics. These assess visual quality and learned image preference alongside the task metrics. QJudge (Li et al., 2026a) averages Quality, Aesthetics, and Alignment, with Creative Generation added for OCR. PickScore (Kirstain et al., 2023) and HPSv3 (Ma et al., 2025) measure prompt-conditioned image preference, while Aesthetic Score (Aesth.) evaluates image-only aesthetics (Schuhmann, 2022). All auxiliary metrics use every image in the test bank; QJudge composites are compared within tasks. HPSv3 is also a training reward for the multi-reward NFT baseline, rather than an independent evaluator for that baseline. Appendix B provides scoring details.

5 Results

Our comparisons ask whether quality gains come from adding a reward, changing the integration of demonstration images, or improving initialization alone. We compare these alternatives, examine cross-backbone results, and then analyze demonstration ratio, reward calibration, and continued supervision after SFT.

5.1 Main comparisons and cross-backbone extensions

Demonstrations improve auxiliary metrics beyond the task reward (Table 1). Across FLUX and SD3.5, Demo-NFT with 5% demonstrations improves QJudge, HPSv3, PickScore, and Aesthetic over single-reward NFT in all four model/task pairs. Task scores improve on FLUX GenEval2

and SD3.5 OCR, with decreases below 0.01 in the remaining comparisons. This pattern supports the target-content distinction in Section 3.3: a single task reward can accompany supervision for multiple visual attributes through the demonstration images.

Adding a preference reward produces a different outcome. The multi-reward NFT baseline obtains higher HPSv3 scores, but lower task scores, QJudge, and PickScore than Demo-NFT in every comparison. Demo-NFT thus offers a favorable task–quality trade-off against this equal-weight reward mixture. Whereas additional rewards reweight native samples, demonstrations also change the image targets (Section 3.3).

With a lower demonstration ratio (5% versus 10%), Demo-NFT achieves higher task scores than DDRL* in three of four settings, while both improve all four auxiliary metrics over NFT. Section 5.2 examines their task–quality trade-offs at a matched demonstration ratio.

Table 1: **Main method comparisons on FLUX and SD3.5.** Demo-NFT uses 5% demonstrations; DDRL* uses 10%. The asterisk denotes our NFT-based adaptation with supervised flow matching on demonstrations. Parentheses specify NFT rewards; task + HPSv3 denotes the multi-reward baseline, which equally averages separately normalized advantages.

Task	Model	Method	Soft-TIFA ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
GenEval2	FLUX	Base	0.6943	51.53	8.44	0.8766	5.57
		NFT (GenEval2)	0.9062	53.90	9.12	0.8807	5.48
		NFT (GenEval2 + HPSv3)	0.8801	57.58	**14.05**	0.8962	5.65
		DDRL*	0.8798	61.37	10.61	**0.9118**	**5.83**
		Demo-NFT	**0.9117**	**61.72**	10.09	0.9031	5.80
	SD3.5	Base	0.6802	47.08	7.84	0.8767	5.52
		NFT (GenEval2)	0.8314	44.21	8.24	0.8752	5.48
		NFT (GenEval2 + HPSv3)	0.8010	43.12	**10.89**	0.8863	5.63
		DDRL*	**0.8364**	54.97	10.16	**0.9091**	**5.77**
		Demo-NFT	0.8291	**56.46**	9.83	0.9027	5.74
Task	Model	Method	OCR score ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
OCR	FLUX	Base	0.6848	57.91	10.84	0.8074	4.89
		NFT (OCR)	**0.8259**	55.73	10.59	0.7944	4.80
		NFT (OCR + HPSv3)	0.7846	56.92	**13.62**	0.8089	5.03
		DDRL*	0.7980	**60.47**	11.78	**0.8355**	**5.27**
		Demo-NFT	0.8188	58.78	11.62	0.8184	5.06
	SD3.5	Base	0.4163	48.90	9.13	0.8162	5.15
		NFT (OCR)	0.5417	37.68	9.14	0.7922	4.84
		NFT (OCR + HPSv3)	0.5502	48.91	**11.53**	0.8359	**5.19**
		DDRL*	0.5400	**51.24**	10.57	0.8332	5.04
		Demo-NFT	**0.5717**	49.90	10.52	**0.8371**	5.07

The benefits extend to other backbones, although their extent varies by task (Table 2). On GenEval2, all four auxiliary metrics improve over NFT for both Qwen and SenseNova, alongside task-score reductions of 1.62 and 0.85 percentage points, respectively. SenseNova OCR improves every reported metric over NFT. Qwen OCR shows a narrower benefit: QJudge improves, but task score, PickScore, and Aesthetic change little, and HPSv3 decreases.

Figure 2 illustrates why text accuracy alone does not characterize a magazine cover. Typography, portrait placement, and visual hierarchy vary across methods alongside remaining text errors, showing aspects of the output beyond OCR reward.

Figure 3 provides corresponding compositional examples, where object presence, color, and a front–back relation must hold together.

5.2 Demonstration ratio and reward capping

Table 3 isolates two design choices around the default Demo-NFT configuration with 5% demonstrations. Removing capping lowers task scores on both GenEval2 settings and on SD3.5 OCR, while leaving FLUX OCR nearly unchanged. On SD3.5 GenEval2, capping raises the score from

Table 2: **Comparisons of Base, NFT, and Demo-NFT.** Results on Qwen-Image and SenseNova for GenEval2 and OCR. NFT and Demo-NFT use matched training setups within each backbone and task.

Task	Model	Method	Soft-TIFA ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
GenEval2	Qwen	Base	0.7940	53.21	**11.65**	**0.9208**	**5.94**
		NFT (GenEval2)	**0.9403**	60.64	10.75	0.9022	5.55
		Demo-NFT	0.9241	**63.37**	10.85	0.9122	5.78
	SenseNova	Base	0.8348	**59.08**	10.00	**0.9230**	**6.21**
		NFT (GenEval2)	**0.9312**	55.76	8.67	0.9010	5.67
		Demo-NFT	0.9227	58.05	**11.47**	0.9225	5.89
Task	Model	Method	OCR score ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
OCR	Qwen	Base	0.8713	68.63	**12.73**	**0.8476**	**5.43**
		NFT (OCR)	0.9730	70.58	12.64	0.8382	5.28
		Demo-NFT	**0.9738**	**72.16**	12.41	0.8381	5.29
	SenseNova	Base	0.9378	76.14	**12.40**	**0.8359**	5.29
		NFT (OCR)	0.9619	71.43	11.85	0.8141	5.12
		Demo-NFT	**0.9761**	**78.13**	12.15	0.8315	**5.42**



Figure 2: **Matched LongText-Extend examples on FLUX and SD3.5.** Columns show Base, NFT, Demo-NFT (w/o cap), and Demo-NFT. The magazine-cover prompt specifies text, typography, and layout. All panels share the prompt and seed, with 50 sampling steps and CFG 3.5.

0.8078 to 0.8291, close to NFT's 0.8314, while retaining gains in all four auxiliary metrics over NFT.

Doubling the demonstration ratio from 5% to 10% yields no consistent gains: GenEval2 task scores decrease on both backbones, while OCR scores improve only marginally. These results support 5% as an effective default that limits demonstration use. Appendix G provides all four ratio/capping combinations, including the 10% variant without capping.

Matching the ratio controls the demonstration allocation when comparing integration strategies. At 10%, Demo-NFT has higher task scores than DDRL* in three of four comparisons, including both FLUX tasks, but DDRL* retains higher HPSv3, PickScore, and Aesthetic on FLUX. On SD3.5, all five metrics favor Demo-NFT for OCR and DDRL* for GenEval2. Thus, the same demonstration source supports different task–quality operating points depending on how it enters the training objective.

Image content and reward calibration. Figure 4 connects image targets, advantages, and gradients on SD3.5 OCR. On 32 matched prompts, demonstrations have higher QJudge Quality and Aesthetics than native samples (a); these gains persist under common task-reward weighting (Appendix F). Holding all 2,048 targets fixed, capping reduces demonstration mean advantage from 3.64 to 0.69



Figure 3: **Matched GenEval2 examples on FLUX and SD3.5, with columns as in Figure 2.** Prompt: "A scene with a candle, a blue truck, and a cookie. The truck is positioned in front of the cookie." All panels share the seed, 50 sampling steps, and CFG 3.5.

Table 3: **Ablating reward capping and demonstration ratio.** Each block starts from Demo-NFT with 5% demonstrations, then removes capping or increases the ratio to 10%. The demonstration fraction is denoted by ρ. Table 6 in Appendix G reports the complete ratio/capping comparison.

Task	Model	Method	ρ	Soft-TIFA ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
GenEval2	FLUX	Demo-NFT	5%	**0.9117**	61.72	**10.09**	0.9031	**5.80**
		Demo-NFT (w/o cap)	5%	0.9029	**63.70**	9.80	0.9006	5.76
		Demo-NFT	10%	0.8944	63.42	10.00	**0.9040**	**5.80**
	SD3.5	Demo-NFT	5%	**0.8291**	**56.46**	**9.83**	**0.9027**	**5.74**
		Demo-NFT (w/o cap)	5%	0.8078	55.11	9.73	0.9003	5.71
		Demo-NFT	10%	0.8208	54.36	9.26	0.8924	5.68
Task	Model	Method	ρ	OCR score ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
OCR	FLUX	Demo-NFT	5%	0.8188	58.78	11.62	0.8184	5.06
		Demo-NFT (w/o cap)	5%	0.8193	**60.19**	**12.01**	**0.8315**	**5.25**
		Demo-NFT	10%	**0.8214**	58.75	11.36	0.8150	5.11
	SD3.5	Demo-NFT	5%	0.5717	49.90	10.52	0.8371	5.07
		Demo-NFT (w/o cap)	5%	0.5506	50.80	10.56	**0.8444**	**5.08**
		Demo-NFT	10%	**0.5738**	**51.47**	**10.66**	0.8412	5.06

and median gradient norm from 0.708 to 0.130 (b,c). The native-sample gradient-norm median changes from 0.055 to 0.060, while the negative-advantage fraction falls from 64.1% to 48.3%. Historically, 99.95% of demonstration rewards reach the cap: it calibrates their influence against retained native-sample rewards while retaining their visual content (Appendix E).

5.3 Continued demonstrations after SFT

Table 4 isolates continued demonstration use from initialization on FLUX and SD3.5. Both branches first learn from the same demonstrations through SFT; only Demo-NFT continues using demonstration targets during RL.

Continued demonstrations improve all four auxiliary metrics in every model/task pair after shared SFT initialization. These gains distinguish their ongoing role during RL from the effect of the starting adapter. On OCR, task performance remains similar for FLUX and improves for SD3.5. On GenEval2, task scores instead fall by 0.0344 and 0.0228, respectively, identifying a quality-oriented trade-off rather than uniform task improvement. Corresponding checkpoint evaluation curves are provided in Appendix B (Figure 5).

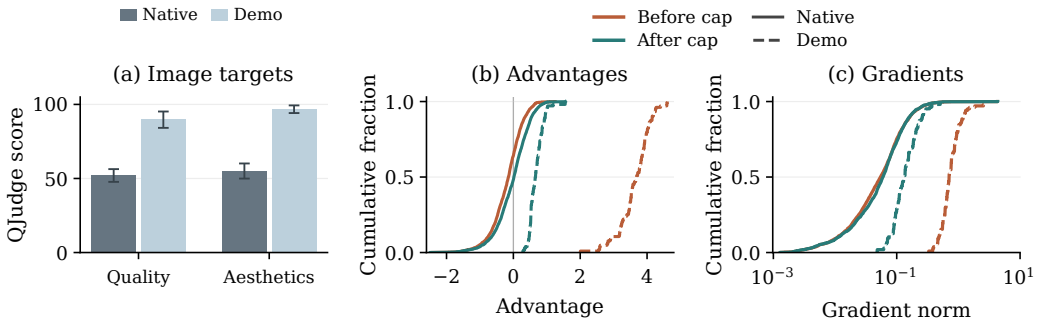


Figure 4: **Image targets, advantages, and gradients.** SD3.5 OCR at checkpoint 360: (a) QJudge source means on 128 targets from 32 matched prompts, with 95% prompt-bootstrap intervals; (b,c) cap effects on advantages and per-target LoRA gradient norms for all 2,048 targets. Solid/dashed lines denote native samples/demonstrations; orange/teal denote before/after cap. Protocols are in Appendices E and F.

Table 4: **Continued demonstrations after shared SFT initialization.** Each pair starts from the same SFT120 adapter and receives up to 360 additional RL updates with native-only NFT or Demo-NFT at 5%.

Task	Model	Method	Soft-TIFA ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
GenEval2	FLUX	NFT (GenEval2)	**0.9348**	60.70	8.39	0.8872	5.67
		Demo-NFT	0.9004	**64.44**	**10.06**	**0.9053**	**5.81**
	SD3.5	NFT (GenEval2)	**0.8685**	50.89	9.22	0.8875	5.62
		Demo-NFT	0.8457	**54.61**	**10.01**	**0.9033**	**5.72**
Task	Model	Method	OCR score ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
OCR	FLUX	NFT (OCR)	**0.8658**	57.80	11.00	0.8032	4.93
		Demo-NFT	0.8614	**59.37**	**11.69**	**0.8297**	**5.16**
	SD3.5	NFT (OCR)	0.5683	38.85	8.81	0.7877	4.62
		Demo-NFT	**0.5754**	**51.06**	**10.61**	**0.8377**	**5.08**

6 Limitations

Results use one training run per configuration. Selecting the best evaluated test checkpoint is descriptive and does not estimate performance after independent validation-based model selection. Visual evaluation relies on automated judges and learned preference models rather than direct human assessment. The cross-backbone extensions use matched within-model comparisons but differ in initialization and demonstration ratio; equal update counts do not establish equal training cost.

7 Discussion

Demo-NFT shows how external image supervision and task-reward feedback can be made complementary within diffusion RL. Demonstrations supply coherent visual targets in which task execution, typography, and layout are realized together. Joint advantage estimation connects learning from these targets to task performance, while native-referenced capping calibrates their influence relative to native samples. This coordination enables a single task reward to guide learning from richer image-level supervision.

Demo-NFT also makes efficient use of demonstration data: a 5% target fraction improves all four auxiliary metrics over NFT across the four main settings, while most training targets remain native samples. Doubling this fraction yields no consistent additional gains. The benefits persist after shared SFT initialization, highlighting continued, reward-guided use of demonstrations rather than initialization alone. Together, these findings support a data-efficient approach to diffusion post-training that coordinates holistic image supervision with task-reward feedback.

AI USE STATEMENT

Generative AI tools were used to produce the synthetic demonstration images and to assist with code development, experimental analysis, and manuscript revision.

REPRODUCIBILITY STATEMENT

Section 3 specifies the training objective and reward calibration, and Section 4 defines the data, comparisons, evaluation metrics, and checkpoint selection. Appendix A gives implementation details. Appendices B and C describe evaluation accounting and the evidence scripts that regenerate and independently audit the numerical tables from stored outputs.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. Official blog post, 2026. URL https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence. January 15, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026. URL https://arxiv.org/abs/2609.11929.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. URL https://arxiv.org/abs/2403.03206.

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. *arXiv preprint arXiv:2305.16381*, 2023.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-Omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. URL https://arxiv.org/abs/2507.22058.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. URL https://arxiv.org/abs/2310.11513v1.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023. URL https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023. URL https://arxiv.org/abs/2307.06350v1.

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.

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. *arXiv preprint arXiv:2305.01569*, 2023. URL https://arxiv.org/abs/2305.01569.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. URL https://arxiv.org/abs/2302.12192.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026a.

Xiaomin Li, Qian Liang, Yinan Li, Ying Zhang, Chen Li, Jing Lyu, Huchuan Lu, and Xu Jia. PortraitGen: Exemplar-driven GRPO with dual-reward guidance for photorealistic portrait generation. *arXiv preprint arXiv:2606.26930*, 2026b. URL https://arxiv.org/abs/2606.26930.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

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. *arXiv preprint arXiv:2505.05470*, 2025. URL https://arxiv.org/abs/2505.05470.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. URL https://arxiv.org/abs/2209.03003.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101.

Yuhang Ma et al. HPSv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. URL https://arxiv.org/abs/2508.03789.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Christoph Schuhmann. LAION-Aesthetics. Official blog post and aesthetic-predictor implementation, 2022. URL https://laion.ai/blog/laion-aesthetics/. August 16, 2022.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wallace_Diffusion_Model_Alignment_Using_Direct_Preference_Optimization_CVPR_2024_paper.html.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

Haotian Ye, Kaiwen Zheng, Jiashu Xu, Puheng Li, Huayu Chen, Jiaqi Han, Sheng Liu, Qinsheng Zhang, Hanzi Mao, Zekun Hao, Prithvijit Chattopadhyay, Dinghao Yang, Liang Feng, Maosheng Liao, Junjie Bai, Ming-Yu Liu, James Zou, and Stefano Ermon. Data-regularized reinforcement learning for diffusion models at scale. *arXiv preprint arXiv:2512.04332*, 2025. URL https://arxiv.org/abs/2512.04332.

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. *arXiv preprint arXiv:2509.16117*, 2025.

A Implementation Details

Clean-state and velocity-space equivalence. NFT's velocity branches are $v_\theta^+ = (1-\beta)v_{\text{old}} + \beta v_\theta$ and $v_\theta^- = (1+\beta)v_{\text{old}} - \beta v_\theta$. Applying $\widehat{z} = z_t - tv$ gives

$$z_t - tv_\theta^+ = (1-\beta)(z_t - tv_{\text{old}}) + \beta(z_t - tv_\theta) = \widehat{z}_\theta^+,$$
$$z_t - tv_\theta^- = (1+\beta)(z_t - tv_{\text{old}}) - \beta(z_t - tv_\theta) = \widehat{z}_\theta^-.$$

Each branch's coefficients sum to one, so the affine conversion preserves the mixed prediction. Under the same reconstruction loss, advantages, and stop-gradient conventions for the reference predictions and loss normalizers, the two expressions therefore give identical losses and trainable gradients. Both branches reconstruct the same clean target; the negative branch does not introduce a separate negative image target.

Demonstration allocation. A deterministic prompt-stratified schedule allocates integer demonstration slots across the distributed batch to realize the nominal replacement ratio ρ. All native samples are generated before replacement, so the model's sampling workload remains fixed across ratios. The full demonstration set remains available for sampling at every ratio.

The main FLUX/SD3.5 NFT recipes use LoRA optimization (Hu et al., 2021) with learning rate 3×10^{-4}, interpolation coefficient $\beta = 0.1$, reference-prediction penalty coefficient 0.02, advantage standard-deviation floor 0.1, advantage clipping to $[-5, 5]$, and gradient-norm clipping at 0.15. The moving-reference decay is $d(s) = \min(0.001s, 0.5)$, updated at outer-epoch boundaries. The main recipes use 16 native samples per prompt and two optimizer updates per outer epoch. Common random numbers share loss timesteps and noise within each prompt group.

Training rollouts use 10 steps without classifier-free guidance; the reported evaluation uses 50 steps with CFG 3.5. Backend sampling conventions are retained: FLUX/SD3.5 use their evaluation pipelines, Qwen uses true classifier-free guidance with an empty negative prompt and its native norm rescaling, and SenseNova uses its native shifted Euler sampler with shift 3 and no CFG norm rescaling. Model-conditioning implementations remain backend-specific.

For DDRL*, demonstration slots use supervised flow matching rather than a constant NFT advantage. For NFT with task and HPSv3 rewards, each reward uses the original NFT advantage rule; the task and preference advantages are averaged with equal weights before the existing NFT clipping/loss, without a second normalization of the mixture. SFT-initialized runs load the adapter from update 120, reset optimizer state and the local RL update counter, and train each branch independently. Cumulative update counts include the 120 initialization updates. Equal update budgets do not imply equal compute or numbers of demonstration presentations.

For SenseNova GenEval2, both methods share a supervised initialization. Demo-NFT uses a small fraction of demonstration targets on both GenEval2 and OCR.

Table 5. Main-backbone NFT configuration. Values are recovered from stored run configurations and adapter metadata for both tasks. These are RL settings. Historical in-training evaluation settings are distinct from the common 50-step, CFG-3.5 protocol used for all reported tables.

Setting	FLUX 4B	SD3.5 Medium
Resolution / mixed precision	1024 / BF16	1024 / BF16
GPUs per run / training seed	16 / 42	16 / 42
LoRA rank / alpha	32 / 64	32 / 64
Optimizer / learning rate (Loshchilov & Hutter, 2017)	AdamW / 3×10^{-4}	AdamW / 3×10^{-4}
Adam moments / epsilon	(0.9, 0.999) / 10^{-8}	(0.9, 0.999) / 10^{-8}
Weight decay	10^{-4}	10^{-4}
Rollout solver / steps / guidance	DPM2 / 10 / 1.0 (no CFG)	DPM2 / 10 / 1.0 (no CFG)
Prompt draws per rollout / targets per draw	32 / 16	32 / 16
Per-GPU rollout batch / batches	8 / 4	2 / 16
Per-GPU training batch / accumulation	2 / 8	1 / 16
Global targets per optimizer update	256	256
Optimizer updates per rollout	2	2
NFT interpolation β	0.1	0.1
Reference-prediction penalty	0.02	0.02

Setting	FLUX 4B	SD3.5 Medium
Advantage normalization	Prompt mean, batch std.	Prompt mean, batch std.
Std. floor / advantage clip	0.1 / $[-5, 5]$	0.1 / $[-5, 5]$
Gradient-norm clip	0.15	0.15
Loss timestep fraction	0.99 (9 timestep draws)	0.99 (9 timestep draws)
Shared loss noise / timestep within prompt	Yes / yes	Yes / yes
Moving-reference decay	$\min(0.001s, 0.5)$	$\min(0.001s, 0.5)$
Gradient checkpointing	No	Yes

The accumulation settings produce the same global target batch despite different per-device micro-batches. Reference decay is updated at outer-epoch boundaries using the optimizer-update counter s. Demonstration ratio and reward treatment vary as specified in each comparison; they do not change the total target count. Pretrained-initialized runs are compared over up to 480 updates, and shared-SFT branches over 360 local RL updates after SFT120. The configuration audit retains the recorded model paths, adapter target modules, and source hashes.

B Evaluation Accounting and Supplementary Metrics

The evidence records 204 generated 50-step, CFG-3.5 evaluation banks, including 105 GenEval2 and 99 LongText-Extend (OCR) banks. GenEval2 uses 128 images per bank and LongText-Extend uses 100. All OCR QJudge results in Tables 1–4 use the complete 100-image bank: 24 distinct selected checkpoints and 2,400 image-level judgments. Including Table 6 brings the total to 26 distinct OCR checkpoints and 2,600 judgments. GenEval2 QJudge uses all 128 images. PickScore, HPSv3, and Aesthetic evaluate every image in each bank. The HPSv3 and Aesthetic sweeps each cover 23,340 images across all 204 banks. Side scores are computed from the existing generated images without regeneration. Explicit N/A criteria are omitted from the corresponding QJudge mean.

Individual QJudge dimensions at the selected checkpoints are tabulated in the accompanying supplementary evidence file, using the same checkpoint selections as the main tables.

Metric implementations. GenEval2 scoring uses Qwen3-VL-8B-Instruct to estimate correct-answer probabilities for prompt-specific questions. Soft-TIFA AM averages these probabilities first within each image and then across images, giving each image equal weight. This reporting aggregation does not change the stored training rewards (Appendix D). The OCR score weights character-similarity-weighted word recall and reading order by 0.8 and 0.2. PickScore reports the implementation's scaled text–image logit divided by 26, not a pairwise win probability.

For the main backbones, original-model runs are selected from updates $\{120, 240, 360, 480\}$, and SFT-initialized branches from local RL updates $\{120, 240, 360\}$. Qwen GenEval2 has evaluations at 120, 240, and 360 for both methods. Qwen OCR additionally has a capped evaluation at 480. SenseNova GenEval2 branches each use the same three local RL checkpoints; its pretrained-initialized OCR runs have evaluations at 120, 240, 360, and 480. Exact ties in task-based selection use the earlier checkpoint.

The full data manifest records source directories, checkpoint identities, generated-image hashes, evaluation settings, and the selected checkpoint for every table row. Reported auxiliary metrics are never independently maximized.

Preference-scoring implementations. HPSv3 reports the mean raw preference prediction for each image and its full prompt. Aesthetic Score (Aesth.) uses the image-only LAION improved aesthetic predictor with OpenAI CLIP ViT-L/14 (Schuhmann, 2022; Radford et al., 2021); we report its raw prediction. Both remain separate from QJudge.

PickScore text truncation. The scorer uses its CLIP tokenizer (Radford et al., 2021) with a 77-token limit, including special tokens, and right truncation. All 100 OCR prompts exceed this limit: their untruncated lengths range from 190 to 285 tokens, with median 238.5. PickScore therefore conditions on a prompt prefix and cannot assess requirements appearing only in the omitted suffix, including some requested text and typography. The same truncation is applied across methods, so the scores remain a matched prefix-conditioned preference comparison, not a measure of full-

prompt compliance. OCR score evaluates the requested text, and QJudge receives the full prompt. In contrast, all 128 GenEval2 prompts fit within the limit (maximum 44 tokens). The tokenizer audit records each prompt's original and retained lengths; no prompts or existing scores are changed by this audit.

Shared-SFT checkpoint evaluations. Figure 5 complements Table 4 with task scores and PickScore at three evaluated checkpoints after the common SFT120 initialization.

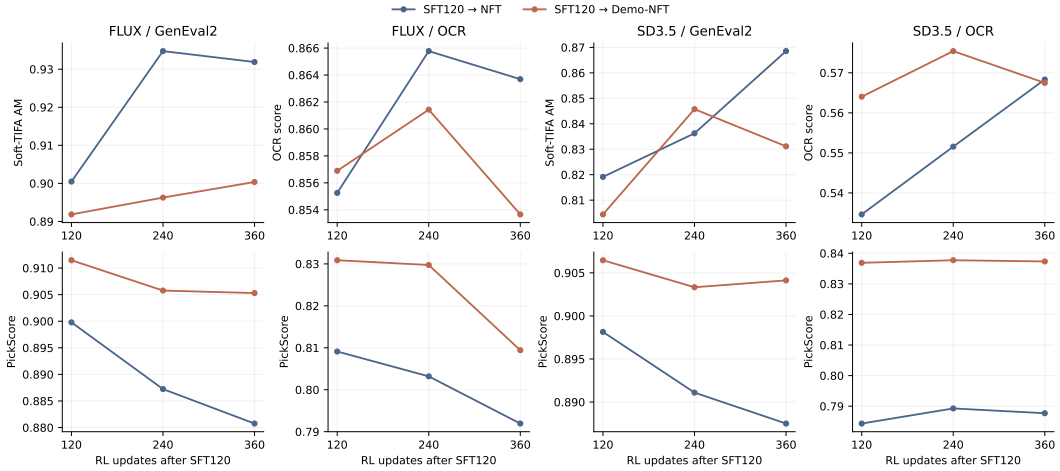


Figure 5: **Shared-SFT checkpoint evaluation curves on the main backbones.** Columns correspond to the two tasks on FLUX and SD3.5; the top row reports task scores and the bottom row reports PickScore. The horizontal axis counts RL updates after the common SFT120 initialization. Every point uses the same 50-step, CFG-3.5 evaluation protocol; no separate checkpoint selection is applied to the preference curves.

C Reproducibility and Data Statement

GenEval2 data preparation. The original collection contains four candidates for each of 2,000 prompts. Low-scoring candidates were regenerated, prompts without an acceptable candidate were removed, and four candidates were retained per remaining prompt, producing the training set internally designated clean-v3. We use a 128-prompt GenEval2 evaluation set with semantics-preserving normalization to the training template. This standardizes scene framing and entity wording without adding visual requirements. Dataset audits check both original and normalized prompt text and semantic signatures against the training sets.

LongText-Extend construction. We adapted and expanded the 160 English prompts in LongText-Bench (Geng et al., 2025) into a 200-prompt reference collection, then used it as few-shot examples for Gemini 3.5 Flash to generate new prompts across 15 scene categories. The generation instructions match the reference style, length, and text density while varying scene content, quoted text, layout, and typography. After removing stray Markdown backticks, the archived generations account for all 2,100 LongText-Extend prompts. The frozen split contains 2,000 training prompts and 100 evaluation prompts, with no exact-text overlap between them; neither split contains exact copies of the reference collection or original benchmark prompts. Per-prompt provenance records identify the source batches and final split.

The numerical tables are generated from stored per-image results by build_uniform50_evidence.py. The script checks image counts, task-summary agreement, prompt pairing, QJudge and PickScore image hashes, and sampling settings. The uniform50_evidence.json artifact stores the extracted results, per-checkpoint scoring counts, and source hashes; uniform50_results.csv records the per-checkpoint metrics, and UNIFORM50_SUPPLEMENT.md contains the selected-checkpoint diagnostic dimensions. The selected OCR QJudge results are frozen in a separately audited 100-image snapshot. The earlier evaluation snapshot is retained separately and is not mixed into the current tables.

The demonstrations are synthetic outputs from GPT-image-2. Data use and release are subject to the applicable model and service terms. The normalized prompts in our 128-prompt GenEval2 evaluation set are audited against the training collections for exact-text and semantic-signature overlap. LongText-Extend uses the generated evaluation split described above. Evaluation prompts have been inspected during development. AI tools assisted manuscript drafting, evidence extraction, and numerical consistency checks.

D Checkpoint Selection and Training Rewards

Checkpoint selection. Main FLUX/SD3.5 tables select the largest evaluated task score for each method, breaking exact ties in favor of the earlier checkpoint. The GenEval2 extensions use the same AM-based rule. The OCR extensions instead compare NFT and capped 5% at update 240 for Qwen and update 360 for SenseNova. These two common-step comparisons were chosen after inspecting the trajectories; they are descriptive comparisons, not pre-specified or validation-selected operating points. All metrics in a row come from the same checkpoint and generated-image bank. The accompanying evidence records every evaluated checkpoint, including those not selected for the tables.

Restored checkpoints. In the SenseNova OCR capped-5% run, the checks at updates 240 and 360 both restored the retained best state from update 200. Checkpoint-360 therefore denotes the state saved after 360 executed updates, including rollback, rather than 360 uninterrupted updates to the retained weights. Its adapter is identical to checkpoint-240, and their fixed-seed evaluation images are identical. The two saves do not constitute independent evidence of improvement. This distinction applies to the OCR extension and should not be confused with the separately trained GenEval2 branches.

Training rewards. For GenEval2, the stored training runs use geometric aggregation of the question-level correct-answer probabilities:

$$r_{\text{train}}(x,p) = \exp\left[\frac{1}{m_p}\sum_{j=1}^{m_p}\log\max\{s_j(x,p), 10^{-12}\}\right].$$

This reward penalizes a low-scoring required attribute more strongly than an arithmetic average. The tables report the Soft-TIFA AM defined in Section 4.4; changing the reporting aggregation does not change the training reward used for those checkpoints. For OCR, both training and task evaluation use $r_{\text{OCR}} = 0.8R_{\text{soft}} + 0.2O$, where R_{soft} is word-level character-similarity recall and O measures reading order. OCR precision is logged but has zero weight. Tokenization ignores capitalization and punctuation, so OCR score is not an exact whole-image text-success rate. DemoNFT applies Equation (1) to these task rewards before joint advantage estimation; reported task scores are uncapped.

Data-regularization control. DDRL* uses ordinary supervised flow matching on demonstration slots and NFT on native slots, averaging their per-target losses over the mixed batch. With demonstration fraction ρ, this gives $(1-\rho)\overline{\mathcal{L}}_{\text{NFT,native}} + \rho\overline{\mathcal{L}}_{\text{FM,demo}}$, with unit per-target FM weight. The asterisk identifies the NFT-based adaptation of DDRL's data-regularization strategy under our shared data protocol. Demonstrations do not enter the native-sample advantage statistics, and their supervised losses are not weighted by task rewards. Matching ρ controls the target allocation, not the magnitudes of the two losses or their gradients.

E Reward-Calibration Diagnostics

Advantage ceiling. Within a prompt, subtracting the shared baseline and dividing by the same positive denominator preserves the capped reward ordering, yielding Equation (5). Subsequent advantage clipping is monotone and preserves this ceiling. Demonstrations at or above the highest retained native reward receive its capped advantage; those below retain their original rewards. Distinct image targets can still produce different reconstruction gradients at the same advantage.

Baseline-shift derivation. On the same mixed batch, let $K = |\mathcal{G}_p|$ and let $\Delta_p = \sum_{d\in\mathcal{D}_p^{\text{used}}}(r_d - \widetilde{r}_d) \geq 0$ be the total demonstration reward reduction for prompt p. Since native-sample rewards are unchanged, $\mu_p^{\text{cap}} = \mu_p^{\text{uncap}} - \Delta_p/K$. Every retained native target therefore satisfies $r_i - \mu_p^{\text{cap}} =$

$(r_i - \mu_p^{\text{uncap}}) + \Delta_p/K$. The normalization denominators in Equation (2) are positive, including when the standard-deviation floor is active. A nonnegative native-sample advantage cannot become negative, and the number of negative native-sample advantages cannot increase. Symmetric clipping to $[-5, 5]$ preserves these signs. However, the batch standard deviation is recomputed, so this does not imply that every numerical native-sample advantage increases.

Gradient-weighting derivation. At fixed model and reference parameters, consider the same mixed batch, loss timesteps, and noise. Let $g_i^{\pm} = \nabla_\theta \ell(\widehat{z}_{\theta,i}^{\pm}, z_i)$ be the branch gradients, and let g_i denote the gradient of the per-target term in Equation (4). With advantages treated as fixed weights and $\bar{A}_i = \text{clip}(A_i, -a, a)$, differentiating Equation (4) gives $g_i = (a/\beta)[q_i g_i^+ + (1-q_i) g_i^-]$. The branch gradients, including the detached reconstruction-loss normalizers, are unchanged between the paired calculations. Subtracting the two gradients and using $q_i^{\text{cap}} - q_i^{\text{uncap}} = (\bar{A}_i^{\text{cap}} - \bar{A}_i^{\text{uncap}})/(2a)$ yields

$$g_i^{\text{cap}} - g_i^{\text{uncap}} = \frac{\bar{A}_i^{\text{cap}} - \bar{A}_i^{\text{uncap}}}{2\beta} \left(g_i^+ - g_i^-\right). \tag{6}$$

This identity describes loss gradients before gradient-norm clipping and optimizer updates. The frozen-base prediction penalty is unchanged on the matched inputs, so its gradient cancels in the difference. Capping modifies the branch mixture through the advantages while preserving image targets and reconstruction losses. The resulting gradient magnitude depends on the branch gradients and, for the batch average, their alignment across targets.

Historical reward records. We archive diagnostics from the 1024-resolution, shared-noise/timestep SD3.5 OCR capped-5% run in Table 3. Each of its 240 outer rollouts contains 32 prompt draws with 16 target slots each. We retain one reward diagnostic per rollout, collected before its two parameter updates. Summary reward means and capped fractions are weighted by the number of demonstrations: 99.95% of demonstration rewards are capped, reducing their mean from 0.949 to 0.583. The source manifest records the training-log and provenance hashes; the accompanying CSV preserves unsmoothed values. Figures 4 and 6 instead use the frozen-policy paired diagnostic described below.

Figure 4 plotting conventions. Panel (a) compares target quality on a 128-image subset, with prompt-level means and bootstrap intervals (Appendix F). Panels (b,c) show empirical cumulative distributions at checkpoint 360, using solid lines for native samples and dashed lines for demonstrations. Orange and teal denote uncapped and capped calculations; panel (c) uses a logarithmic horizontal axis for per-target gradient norms. Each source's cumulative distribution is normalized by its own count: 1,944 native samples and 104 demonstrations.

For rollout t, let M_t be the number of demonstration slots, $\bar{\delta}_t$ their logged mean reward reduction, and $K_{p,t}$ the number of targets sharing prompt p. With $N = 512$ targets, the target-weighted mean prompt-baseline reduction is exactly

$$\sum_p \frac{K_{p,t}}{N} \frac{\Delta_{p,t}}{K_{p,t}} = \frac{M_t \bar{\delta}_t}{N}.$$

Each retained native target's advantage numerator increases by its prompt's baseline reduction, while the final normalized advantage also depends on the batch standard deviation. The historical logs record aggregate rewards; the paired diagnostic measures the resulting normalized advantages for individual targets (Figure 6).

Frozen-policy paired diagnostic. We load the stored rollout-policy adapters at checkpoints 120 and 360, rather than the EMA adapters used for the evaluation tables. At each checkpoint, four fresh batches contain 512 targets each: 486 retained native samples and 26 prompt-matched demonstrations, following the original cumulative 5% replacement rule. Sampling uses the training protocol: 1024 resolution, deterministic DPM2 with 10 steps, and no classifier-free guidance. Prompts, generation seeds, replacement positions, and demonstration candidates are matched across checkpoints. This yields 4,096 targets in total.

Within each checkpoint, both calculations use exactly the same images and original OCR rewards. One retains the demonstration rewards; the other applies native-referenced capping. Each then recomputes prompt means and the global batch standard deviation, with floor 0.1, using the training implementation. Prompt grouping follows the trainer's decoded token keys, including repeated prompts across microbatches. We record the normalized advantages, their clipped values in $[-5, 5]$, and the resulting NFT positive-branch mixture coefficients. This is a paired calculation at fixed poli-

cies, not a comparison between independently trained capped and uncapped runs; no parameters are updated.

All 104 demonstrations sampled at each checkpoint are capped. At checkpoint 120, their mean advantage decreases from 3.57 to 0.74; at checkpoint 360, it decreases from 3.64 to 0.69. The native-sample negative-advantage fraction falls from 62.2% to 48.2% and from 64.1% to 48.3%, respectively (reductions of 14.0 and 15.9 percentage points before rounding). No sample reaches the advantage-clipping threshold, so these differences remain in the final branch weights. The mean demonstration positive-branch coefficient decreases from 0.8565 to 0.5735 at checkpoint 120 and from 0.8636 to 0.5686 at checkpoint 360; the corresponding native-sample means increase from 0.4809 to 0.4961 and from 0.4805 to 0.4963. These coefficients quantify the positive-branch weighting in the NFT loss. The four batches share a trained checkpoint and are not independent training seeds.

Paired gradient diagnostic. At checkpoint 360, we deterministically VAE-encode the 2,048 diagnostic images and evaluate the NFT loss using the saved current and old-policy adapters. Each rollout batch uses nine stratified scheduler timesteps, with Gaussian noise shared among targets of the same full prompt. Both reward treatments use identical encoded targets, parameters, timesteps, and noise. We measure the L2 norm of each target's LoRA gradient after averaging its NFT loss over these timesteps. Advantages and loss normalizers are detached, as in training. Measurement precedes gradient clipping and excludes the reference penalty; no optimizer step is taken.

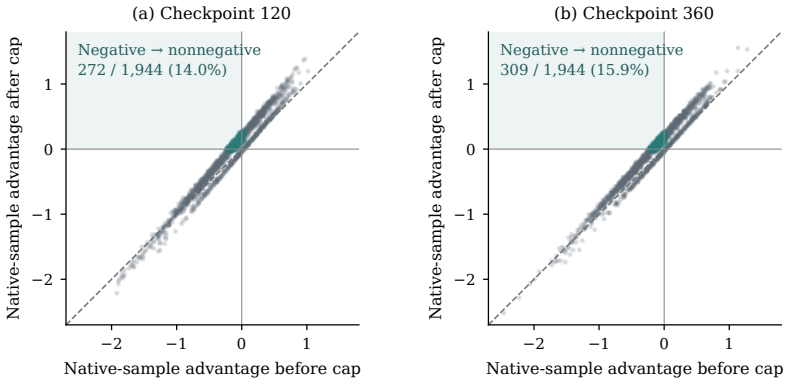


Figure 6: **Paired native-sample advantage changes at checkpoints (a) 120 and (b) 360.** Each point is the same native target before and after capping; both panels share axis limits. The dashed diagonal denotes no change. Teal points in the shaded upper-left quadrant change from negative to nonnegative advantage: 272 of 1,944 targets (14.0%) at checkpoint 120 and 309 of 1,944 (15.9%) at checkpoint 360. No target changes in the reverse direction.

The paired diagnostic preserves every image, reward, source label, and derived weight. An independent audit reconstructs 49,152 numerical entries from the raw records and verifies that the training prompts and demonstration manifest match the original run.

F Image-Target Supervision Beyond the Task Reward

We analyze image-target content separately from reward calibration. Following DiffusionNFT's reward-weighted data-distribution view (Zheng et al., 2025), we characterize the positive-target distribution supplied to learning.

Reward-weighted target distributions. Fix a prompt p and suppress conditioning on it. Let π_N and π_D denote distributions of retained native and demonstration image targets. Their mixture is $\pi_{\text{mix}} = (1-\rho_p)\pi_N + \rho_p\pi_D$, where $0 < \rho_p < 1$ is the prompt-conditional demonstration fraction, not necessarily the batch-level ratio ρ. Let $w(r) \geq 0$ be the same fixed task-reward weighting function for both sources, with $0 < Z_N = \mathbb{E}_{\pi_N}[w(r)] < \infty$ and $0 < Z_D = \mathbb{E}_{\pi_D}[w(r)] < \infty$. Define $\pi_N^w = w(r)\pi_N/Z_N$ and $\pi_D^w = w(r)\pi_D/Z_D$, and normalize the mixed distribution in the

same way. Then

$$\pi^w_{\text{mix}} = (1-\alpha_p)\pi^w_N + \alpha_p \pi^w_D,$$
$$\alpha_p = \frac{\rho_p Z_D}{(1-\rho_p)Z_N + \rho_p Z_D}. \tag{7}$$

To obtain Equation (7), multiply the original mixture by $w(r)$ and divide by its normalizer $(1-\rho_p)Z_N + \rho_p Z_D$. For $w(r) = r$ with rewards in $[0, 1]$, this is DiffusionNFT's positive-data weighting applied to a mixture of target sources. This calculation uses raw task rewards and fixed source-independent weights to isolate image-target content.

Let $h_j(x, p)$ be an auxiliary image attribute not used as a training reward, with finite absolute expectation under both weighted source distributions. Taking expectations in Equation (7) gives

$$\mathbb{E}_{\pi^w_{\text{mix}}}[h_j] - \mathbb{E}_{\pi^w_N}[h_j] = \alpha_p \left(\mathbb{E}_{\pi^w_D}[h_j] - \mathbb{E}_{\pi^w_N}[h_j]\right). \tag{8}$$

Thus, if reward-weighted demonstrations have higher expected values of several auxiliary attributes than reward-weighted native targets, the mixed positive-target distribution improves on all those attributes. The sufficient condition is higher auxiliary quality after applying the same task-reward weighting. If the task reward is constant and its weight is positive, reweighting leaves each source unchanged; any auxiliary difference in Equation (8) then comes entirely from image content. The attributes need not be explicitly scored during training.

Connection to NFT optimization. Both target sources are optimized with Equation (4), including its positive and negative branches. At fixed model and reference parameters, timesteps, and noise, let $g_i^{\pm} = \nabla_\theta \ell(\widehat{z}^{\pm}_{\theta,i}, z_i)$ and $\bar{A}_i = \text{clip}(A_i, -a, a)$. Treating advantages and reconstruction-loss normalizers as detached, the per-target NFT-loss gradient is

$$g_i = \frac{a}{2\beta}(g_i^+ + g_i^-) + \frac{\bar{A}_i}{2\beta}(g_i^+ - g_i^-). \tag{9}$$

Substituting $q_i = \frac{1}{2} + \bar{A}_i/(2a)$ into $g_i = (a/\beta)[q_i g_i^+ + (1-q_i)g_i^-]$ gives Equation (9). The image target enters both branch gradients through the reconstruction residual and the noised input. Thus, even targets with the same advantage can produce different update directions. For a batch of B targets, their contributions combine as $B^{-1}\sum_i g_i$; the frozen-base prediction penalty is added as specified in Section 3.3, followed by the existing optimizer processing.

Equations (8) and (9) connect target content to the NFT learning signal. Equation (8) identifies the relevant source property: higher auxiliary-attribute values after the same task-reward weighting. Equation (9) shows how those images enter optimization through both the reconstruction residual and the noised input, with advantages modulating the resulting gradients. Demonstrations thereby supply visual properties jointly through complete image targets, rather than encoding each property as an additional scalar reward. Native-referenced capping separately calibrates their relative influence, as analyzed in Section 3.4 and Appendix E.

Paired target-content diagnostic. We score a subset of the checkpoint-360 SD3.5 OCR targets used in Figure 4(b,c): 32 training prompts, each with three native images and one demonstration. Prompts and images are selected by a fixed hash ordering before quality scoring, independently of rewards and quality scores. The native images use the 1024-resolution, 10-step, no-CFG training sampler. We evaluate full prompts with QJudge Quality and Aesthetics separately, and with the same official-compatible HPSv3 implementation used in evaluation. Source means give each prompt equal weight; within a prompt, native-sample scores are averaged over its three images. Intervals use 5,000 paired prompt-bootstrap resamples with seed 42.

Figure 4(a) shows Quality means of 52.33 for native samples and 90.00 for demonstrations, and Aesthetics means of 55.06 and 97.17. HPSv3 means are 9.82 and 12.93. Under the common raw-task-reward weighting of Equation (8), the corresponding native/demonstration means are 52.24/90.00, 55.00/97.17, and 9.81/12.93. The weighted source differences have 95% intervals of [31.61, 44.16], [36.83, 47.26], and [2.85, 3.38], respectively. Thus, this target set satisfies the higher auxiliary-quality condition in Equation (8). Source quality gaps also persist when weighting by the recorded capped NFT positive-branch coefficients. Mean OCR rewards are 0.522 and 0.945: the diagnostic measures the visual content supplied by these two sources, with their observed task-reward difference retained.

G Complete Ratio and Capping Ablation

Table 6 reports both demonstration ratios with and without reward capping, extending the three-row ablation in Table 3. All entries retain the original checkpoint selections and their paired task and auxiliary scores. Without capping, 5% demonstrations obtain higher task scores than 10% on all four model/task pairs. At either ratio, capping improves SD3.5 task scores on both tasks. On FLUX OCR at 10%, capping improves the task score but lowers all four auxiliary scores, illustrating the interaction between demonstration allocation and reward calibration.

Table 6: **Complete demonstration-ratio and reward-capping comparison.** Demo-NFT includes capping; Demo-NFT (w/o cap) removes it. Each model/task block reports both methods at $\rho = 5\%$ and 10%, using the same metric definitions and checkpoint selections as the main tables.

Task	Model	Method	ρ	Soft-TIFA ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
GenEval2	FLUX	Demo-NFT	5%	**0.9117**	61.72	<u>10.09</u>	0.9031	<u>5.80</u>
		Demo-NFT (w/o cap)	5%	<u>0.9029</u>	**63.70**	9.80	0.9006	5.76
		Demo-NFT	10%	0.8944	<u>63.42</u>	10.00	<u>0.9040</u>	<u>5.80</u>
		Demo-NFT (w/o cap)	10%	0.8945	62.32	**10.36**	**0.9075**	**5.83**
	SD3.5	Demo-NFT	5%	**0.8291**	**56.46**	**9.83**	**0.9027**	**5.74**
		Demo-NFT (w/o cap)	5%	0.8078	<u>55.11</u>	<u>9.73</u>	<u>0.9003</u>	<u>5.71</u>
		Demo-NFT	10%	<u>0.8208</u>	54.36	9.26	0.8924	5.68
		Demo-NFT (w/o cap)	10%	0.8007	53.58	9.65	0.8974	5.68
Task	Model	Method	ρ	OCR score ↑	QJudge ↑	HPSv3 ↑	Pick ↑	Aesth. ↑
OCR	FLUX	Demo-NFT	5%	0.8188	58.78	11.62	0.8184	5.06
		Demo-NFT (w/o cap)	5%	<u>0.8193</u>	<u>60.19</u>	**12.01**	<u>0.8315</u>	<u>5.25</u>
		Demo-NFT	10%	**0.8214**	58.75	11.36	0.8150	5.11
		Demo-NFT (w/o cap)	10%	0.8036	**60.81**	<u>11.95</u>	**0.8329**	**5.27**
	SD3.5	Demo-NFT	5%	<u>0.5717</u>	49.90	10.52	0.8371	<u>5.07</u>
		Demo-NFT (w/o cap)	5%	0.5506	50.80	10.56	**0.8444**	**5.08**
		Demo-NFT	10%	**0.5738**	**51.47**	**10.66**	<u>0.8412</u>	5.06
		Demo-NFT (w/o cap)	10%	0.5390	<u>51.07</u>	<u>10.62</u>	0.8389	5.05

H Training and Test Prompt Examples

We list ten training and ten test prompts for each task, sampled uniformly without replacement from each split with random seed 42. Prompts are reproduced in full as used in the experiments; the GenEval2 test examples use the normalized evaluation wording.

H.1 GenEval2: training prompts

Training 01. A scene with two donuts, three giraffes, and two birds. The donuts are positioned in front of the giraffes. The giraffes are chasing the birds.

Training 02. A scene with four dogs, two bicycles, and two horses. The dogs are positioned to the left of the bicycles. The bicycles are positioned directly above the horses.

Training 03. A scene with five candles and three checkered sheep. The candles are positioned in front of the sheep.

Training 04. A scene with four white stone koalas and three rabbits. The koalas are jumping over the rabbits.

Training 05. A scene with a guitar and a pink suitcase.

Training 06. A scene with a checkered sheep and two striped yellow turtles. The sheep is positioned in front of the turtles.

Training 07. A scene with four croissants and four umbrellas.

Training 08. A scene with five trucks and four metal raccoons. The trucks are positioned directly below the raccoons.

Training 09. A scene with four striped candles.

Training 10. A scene with four striped monkeys.

H.2 GENEVAL2: TEST PROMPTS

Test 01. A scene with two blue cars.

Test 02. A scene with a striped kangaroo and a rabbit.

Test 03. A scene with a stone suitcase and a trumpet. The suitcase is positioned behind the trumpet.

Test 04. A scene with three motorcycles and a brown giraffe.

Test 05. A scene with five backpacks and four black suitcases.

Test 06. A scene with a purple cookie and two sparkling mushrooms.

Test 07. A scene with a penguin, a violin, and a green umbrella. The penguin is positioned directly below the violin. The violin is positioned directly below the umbrella.

Test 08. A scene with two motorcycles, two donuts, and a clock. The donuts are positioned in front of the clock.

Test 09. A scene with an elephant, three red umbrellas, and two metal motorcycles.

Test 10. A scene with seven zebras, seven yellow turtles, and a plastic croissant. The zebras are positioned directly below the turtles. The turtles are positioned to the left of the croissant.

H.3 LONGTEXT-EXTEND (OCR): TRAINING PROMPTS

Training 01. A low-angle, dramatic comic panel capturing a critical negotiation under a massive cave arch, with overcast grey daylight filtering inside. Dame Constance of Ironwood, clad in dark, matte-black plate armor, stands holding a large parchment scroll before Ignis the Great, a massive red dragon with smoking nostrils and heavy gold horns. At the top-left, Dame Constance's speech bubble, drawn in a neat, professional rectangular shape with thin black lines, reads: "We offer a perpetual non-aggression pact, signed by King Aldous, and three hundred barrels of sweet cider delivered annually!". In the upper-right, Ignis's speech bubble, large and rounded with fiery orange accents and bold blackletter font, responds: "Three hundred barrels of cider is a fair start, but I also demand ten barrels of premium honey-mead every midsummer!". On a stone pedestal between them, the title of the document reads "TREATY OF IRONWOOD - PAGE 1" in formal copperplate script. At the bottom-left, a stacked pile of oak casks includes one clearly stenciled with "SWEET CIDER - 50 GAL" and another marked "HONEY-MEAD". Behind them, a piece of cave wall features charcoal graffiti reading "PEACE OR FIRE".

Training 02. An overhead view of a thick, stapled marital settlement agreement lying on a grey metal desk under cool, clinical fluorescent office lighting. The document's first page is displayed, showing a clean, structured layout with standard legal spacing. At the very top, the court jurisdiction is printed in bold, uppercase sans-serif letters: "SUPERIOR COURT OF CALIFORNIA, COUNTY OF ORANGE". Directly below, a boxed section contains the case number: "Case No. 18FL004921". Centered on the page in a heavy, underlined serif font is the document title: "MARITAL SETTLEMENT AGREEMENT". In the middle section, a critical custody clause reads: "The parties agree to joint legal and physical custody". Towards the bottom, a formal signature block is dated: "Signed this 14th day of February, 2020". Underneath the signature lines, a professional stamp reads: "Approved as to form and content". In the bottom-left corner, a round blue ink notary stamp reads: "Notary Public, State of California". A silver paperclip holds several subsequent pages together, peeking out from the top-left corner.

Training 03. A close-up shot focuses on the main directory wall of the reception area at "Apex Spine & Joint Institute", where bright midday sun highlights pale green walls and light oak paneling. The institution's name is proudly displayed at the top of a modern modular glass signage board in clean, brushed aluminum letters reading "APEX SPINE & JOINT INSTITUTE - REHABILITATION WING". Underneath, clearly organized sections indicate physical therapy facilities, listing "Physical Therapy Gym & Hydrotherapy Pools", "Occupational Therapy & Prosthetic Fit-

ting Clinic", and "Pain Management & Sports Medicine Offices" in highly legible sans-serif type. A directional arrow graphic points toward adjacent corridors, accompanying text that reads "<- Main Reception, Pharmacy, and Café Apex". Adjacent to the main glass panel, a blue accessibility plaque features a wheelchair symbol and states "Fully Accessible Restrooms Located Behind Elevator Lobby". Near the automatic glass entry doors, a freestanding sign displays "Masks Recommended for All Patients and Visitors - Thank You" in bold, welcoming lettering.

Training 04. A crisp, high-definition broadcast feed from Stockholm Palace Chapel captures a royal wedding ceremony under the cool, overcast midday light filtering through high arched windows, illuminating a regal color palette of Swedish blue and gold. The camera presents a medium tracking shot of the bride gracefully walking down the aisle. In the upper-right corner of the screen, a glowing blue badge displays "LIVE HD" in white sans-serif lettering, while the top-left corner features a dual-line black bar reading "ROYAL PALACE CHAPEL, STOCKHOLM" and "13:05 CET" in clean, white monospaced font. A small, elegant network logo in the bottom-right corner displays "SVT EXCLUSIVE" in a stylized, hand-lettered gold font. Across the bottom of the screen, a wide, navy-blue caption bar with a thin gold border contains bold, white serif text stating: "Princess Sofia of Sweden enters the chapel escorted by His Majesty King Carl XVI Gustaf". Directly below this, a smaller white sub-caption identifies the prominent guests: "Witnessed by Prime Minister Ulf Kristersson and international heads of state". A thin, light-blue ticker along the very bottom edge reads: "Up next: The official balcony appearance and flypast over Stockholm Palace".

Training 05. An eye-level medium close-up of a lineup of four craft beer cans arranged on a white marble kitchen countertop, bathed in soft, natural morning light filtering through a nearby window. The cans feature eco-friendly matte paper labels with watercolor forest illustrations in green and gold tones. The top-center of each label displays the brand mark "GREEN FIELD BREWING" in an elegant, embossed copperplate serif font. The primary can in focus is labeled "PACIFIC DRIFT IPA" in a crisp, modern sans-serif typeface, immediately above the technical stats "6.2% ABV | 55 IBU" and a descriptive text block reading "Hop Profile: Galaxy, Nelson Sauvin, and Motueka". Next to it, another can is titled "HARVEST MOON WHEAT" in a flowing hand-written script. The side of the furthest can clearly shows the brewery's address "902 Coastal Highway, Santa Cruz, CA 95060" printed in clean grey lettering, underneath a small green leaf icon with the text "Sustainably Brewed in California" and "100% Organic Barley". Tiny water droplets are beaded across the textured labels, giving them a fresh, chilled look.

Training 06. An overcast afternoon television broadcast of a women's championship football final, with a medium close-up of a female manager in a dark rain jacket shouting instructions from the technical area. In the top-left corner, a sleek purple-and-silver scoreboard overlay reads "CAS 1 - 1 ESW" with the digital clock showing "53:14". Just beneath this, a purple banner displays "CHAMPIONSHIP FINAL" in elegant bold serif lettering. A horizontal stat overlay at the bottom left reads "TOTAL ATTEMPTS: CASCADIA FC 14 | EMPIRE STATE WOMEN 10". The lower-right corner of the broadcast features a bright yellow sponsor logo for "APEX ATHLETICS". Centered at the bottom, a semi-transparent dark subtitle bar presents the live commentator caption: "Live Feed: Manager Sarah Jenkins is urgently gesturing for her players to press higher up the pitch to exploit the sudden gaps in the Empire State midfield.". In the upper-right corner, a subtle gray logo reads "GLOBAL TV SPORTS".

Training 07. A night scene illuminated by vibrant neon lights captures a wet cobblestone street lined with maritime businesses at a busy commercial port. The entrance of a brick building features a glowing green neon sign reading "ANCHORAGE CUSTOMS BROKERS". On the frosted glass window below, black gold-leaf lettering states "LICENSED BY THE U.S. CUSTOMS SERVICE - AGENT NO. 9942". Right next door, a hanging wooden sign shaped like an anchor is lit by a warm spotlight, reading "THE PILOT HOUSE TAVERN — COLD BEER & HOT MEALS FOR MARINERS SINCE 1964" in hand-carved, gilded letters. On the sidewalk, a classic wooden A-frame chalkboard sign lists "TONIGHT'S SPECIAL: CLAM CHOWDER & FISH TACOS $14.95" in neat, colorful chalk handwriting. Painted on the brick wall between the two establishments, a bold white warning reads "NO PARKING - ACTIVE SHIP LOADING ZONE 24/7" to keep the driveway clear. A small, historical bronze plaque near the tavern door simply states "FOUNDED 1952" in raised serif letters. Puddles on the asphalt reflect the flashing red, green, and yellow lights of the industrial district.

Training 08. A high-angle photograph of a premium smartphone resting on a dark walnut nightstand at midnight, illuminated by the warm, dim orange glow of a nearby Himalayan salt lamp. The phone displays a night-themed sleep-tracking application with a deep navy and indigo user interface. In a glowing, rounded sans-serif font, the top-center header proudly reads "Somnus: Sleep & Dream Companion for Restful Nights". Below the header, a small gold crescent moon icon is accompanied by the streak counter text "Active Night Streak: 42 Nights of Rest". The primary card on the screen highlights a breathing guide labeled "Wind Down Breathing: 4-7-8 Deep Relaxation Technique", designed to prepare the body for deep sleep. In the audio playlist section underneath, the featured ambient track is titled "Rainfall over Kyoto Bamboo Forest with Soft Flute" with a small credit line beneath that reads "Narrated by Haruto Tanaka, Certified Sleep Guide". At the bottom center, a soft blue glowing call-to-action button is labeled "Begin Sleep Journey Now". The high-contrast, glowing typography stands out sharply against the dark interface, offering a cozy, sleep-inducing aesthetic.

Training 09. A warm, low-angle shot captures a glossy cardboard medicine carton resting on a polished mahogany bedside table, illuminated by the soft golden light of a late afternoon sun. Next to a condensation-beaded glass of water, the amber-tinted box displays highly legible black typography on a clean white adhesive label. The top of the label clearly reads "Lisinopril 20mg - USP — Antihypertensive". Below the drug name, the local pharmacy's header is printed in a friendly, dark blue sans-serif font: "Oakridge Community Pharmacy - 104 Main St. — Tel: 555-0143". The central instruction block features bold, typewriter-style print detailing "Take one tablet daily in the morning for high blood pressure with a full glass of water". A vivid orange auxiliary warning strip affixed to the bottom edge reads "ATTENTION: Avoid salt substitutes containing potassium unless directed by your doctor". On the side flap, printed in small, precise letters, is "Rx 88201-9 — Pharmacist: Sarah Jenkins, PharmD", while the lower-right corner is stamped with "Filled: 05/14/2025 — Exp: 05/2026". The scene exudes a quiet, domestic calm with its rich amber tones and soft shadows.

Training 10. A beautiful afternoon shot of a Canadian Pacific Telegraphs form from September 1945, placed on a rustic cedar table by an open window overlooking a sunny garden. The top of the form features the bold, stenciled header "CANADIAN PACIFIC TELEGRAPHS - WORLD WIDE COMMUNICATIONS". Below this, a printed line reads "CABLE SERVICE TO ALL THE WORLD - WIRELESS TO SHIPS AT SEA". The recipient address is typed clearly: "TO: MRS. EVELYN DUBOIS, 85 RUE ST-JEAN, QUEBEC CITY, QUEBEC". The emotional, typewritten message, with tiny faded letters, reads: "RELEASING FROM SERVICE TOMORROW. BOARDING TRAIN FOR QUEBEC IMMEDIATELY. CANNOT WAIT TO SEE YOU. ALL MY LOVE ALWAYS, PIERRE." A green operator's stamp in the lower-left corner reads "RECEIVED AT QUEBEC TERMINAL - SEP 12 1945". A faint blue hand-stamped slogan at the bottom of the form reads "BUY VICTORY BONDS - FOR CANADA'S FUTURE".

H.4 LongText-Extend (OCR): test prompts

Test 01. A medium close-up shot of a stern, middle-aged male prosecutor interviewed in a dimly lit office at night, surrounded by towering stacks of legal files and cardboard evidence boxes. A single banker's lamp on his desk casts a warm green and golden glow across his face, highlighting his intense expression. In the upper-right, a sleek white copperplate logo reads "JUSTICE ON TRIAL", while the upper-left shows a small white text label indicating "S1:E5 — THE WHISTLEBLOWER". In the lower-left corner, a solid black banner displays his name "Robert Vance, Esq." in bold, white sans-serif, and directly below it, his former role "Assistant District Attorney, 1995-2002" appears in a lighter, gold-colored serif font. Centered at the bottom, a semi-transparent black subtitle bar displays his spoken words in a clean white Arial font: "We knew we only had one shot at this grand jury indictment, so we had to make sure every single document was completely bulletproof." In the lower-right corner, a small, white digital stamp reads "DOCKET #96-CV-410". The contrast between the warm green lamp light and the deep shadows of the office creates a tense, high-stakes legal drama atmosphere.

Test 02. An overhead photograph of an elegant doctoral diploma resting on a polished mahogany desk. The certificate is printed on heavy, off-white cotton parchment paper, featuring an intricate border of embossed gold foil vines. At the top center, the institution's name is rendered in a formal, raised blackletter font: "St. Jude's University of Science". Beneath this, in a smaller, clean serif typeface, the conferring statement reads: "Upon recommendation of the Faculty and by virtue of the

Authority vested in us". The degree itself is displayed prominently in a larger, bold Roman script: "Doctor of Philosophy in Astrophysics". Directly below, the candidate's name is exquisitely hand-lettered in sweeping, copperplate calligraphy, reading: "Dr. Helena Vance Vance-Sterling". A brief paragraph in smaller italics explains the achievement: "In witness whereof we have set our hands and the seal of the University this twelfth day of October, 2021". At the bottom left, a metallic gold seal shines under a soft desk lamp, while the bottom right features two authentic ink signatures with printed titles underneath: "Chancellor Marcus Thorne, PhD" and "Dean of Graduate Studies Sarah Lin".

Test 03. A cinematic, high-contrast screenshot from a 1940s noir detective adventure game, set in a dimly lit, smoky private investigator's office at midnight. Rain streams down the windowpane, casting long, dramatic shadows of Venetian blinds across the scene. A minimalist, dark grey dialogue overlay is positioned at the bottom. In a elegant, white typewriter font, the name of the client is shown as "Evelyn Sterling". Her dialogue text below, in a clean, sans-serif font, reads: "My husband didn't just disappear, Detective. He was investigating the Falcone shipping yards before he went missing. Please, find him." Below her plea, three branching dialogue choices are listed vertically with subtle white underlines. The professional first option reads: "I'll take the case. Where should I start?", the cautious second option states: "This sounds too dangerous for a solo detective.", and the business-minded third option says: "My fee is fifty dollars a day, plus expenses." In the top-right corner of the screen, a subtle HUD element displays "Current Case: Missing Husband", and in the bottom-right corner, a tiny text prompt reads "Press [Space] to Continue", reinforcing the classic, suspenseful hardboiled atmosphere of the detective genre.

Test 04. A low-angle shot of a weathered political campaign poster affixed to a distressed red brick wall in downtown Oakridge, captured during the overcast late-afternoon. The poster features a bold, graphic portrait of a female candidate in duotone navy and cream. At the top of the poster, the candidate's name "ELECT MAYOR EVELYN VANCE" is written in a strong, distressed stencil font. Directly beneath her portrait, the main campaign slogan is displayed in prominent, hand-lettered gold-leaf script: "Building Our Tomorrow, Together". In the middle section, three policy bullets are neatly aligned to the left in a clean, modern sans-serif typeface: "Invest in Public Schools", "Revitalize Downtown District", and "Expand Green Energy Jobs". Near the bottom, printed in a sharp, high-contrast white banner, is the voter registration warning: "Voter Registration Deadline: October 11, 2024". Finally, at the very bottom edge, details for the upcoming rally are laid out in smaller, condensed block lettering: "Join the Rally at Oakridge Park — Oct 15 at 5:00 PM".

Test 05. A close-up, candlelit view of a vintage steamship postcard lying inside the tray of a distressed leather travel trunk. The warm, flickering candlelight illuminates the textured, cream-colored cardstock. The postcard's header reads "Compagnie Générale Transatlantique - French Line" in sophisticated black lithographic print. On the left side, a romantic message written in delicate, hurried French-style cursive ink states "My darling, Paris was lovely, but the sea voyage is divine.". On the right, the destination address is clearly penned as "Monsieur Henri Dubois, Rue de Rivoli 74, Paris, France". A circular, dark purple postmark in the upper-right corner reads "Le Havre Port - 14-8-1909", overlapping a small green stamp labeled "REPUBLIQUE FRANCAISE - 10c". Vertically along the left edge, crisp, tiny lettering reads "Steamship La Provence - Souvenir Post Card". Beside the postcard rests an ornate silver pocket watch showing the time as ten past ten, a fountain pen with a gold nib, and a small pile of dried lavender buds.

Test 06. A clean, modern workspace shot from a straight overhead angle, featuring an open engineering journal centered on a slate-gray desk mat next to a mechanical keyboard and a silver stylus. The journal header at the top-left reads: "IEEE Transactions on Intelligent Systems, Vol. 12, July 2025". The paper's title is printed in a bold, futuristic sans-serif font: "Deep Reinforcement Learning for Autonomous Drone Swarms in GPS-Denied Environments". The dense abstract block begins: "This paper introduces a decentralized swarm-coordination algorithm utilizing multi-agent proximal policy optimization." On the right page, a section heading in medium bold reads: "4.1. Simulation Environment Setup". Underneath a complex scatter plot, a detailed caption reads: "Fig. 7: Simulated trajectories of 12 quadcopters navigating a cluttered 3D obstacle field." A tiny footnote at the bottom of the column is printed in grey text: "This research was funded by the Defense Advanced Research Projects Agency under contract No. HR0011-24-C-0052."

Test 07. A wide, eye-level street-view shot of Plaza 82 during a clear late afternoon, with dramatic, elongated shadows cast by roadside utility poles across the asphalt parking lot. A classic Italian pizza

parlor anchor tenant features a retro red neon sign reading “PAPA’S PIZZERIA - FAMILY RECIPE SINCE 1974” in stylized cursive. Underneath its front window, a bright yellow banner screams “2 LARGE 1-TOPPING PIZZAS FOR ONLY $19.99!” in bold comic-style lettering. Adjacent to the pizzeria, a mobile retailer’s sign reads “SMART WIRELESS - AUTHORIZED DEALER” in a clean, futuristic sans-serif font, with a window poster stating “SWITCH TO US & GET A FREE 5G PHONE TODAY” in energetic, stenciled orange letters. At the corner edge of the plaza, a Metro Gas station displays its prices on a large sign reading “REGULAR $3.59” and “SUPER UNLEADED $4.19” in bright green and red digital displays, while a secondary placard below promises “FREE CAR WASH WITH FILL-UP OF 10 GAL OR MORE” in bold block letters. The low-angled golden sunlight beautifully illuminates the colorful array of commercial text, creating an authentic suburban retail atmosphere.

Test 08. A close-up, angled shot of an interactive digital touchscreen kiosk on a bustling subway platform at dusk, casting a soft glow on its surroundings. The screen displays a dark-mode route map titled “Shin-Yoko Transit Link: Interactive Route Map” in a vibrant, futuristic light-blue font. Bright, glowing traces represent different routes, notably the electric pink path labeled “Sakura Express (Pink Line)” and the intersecting light-blue path labeled “Skyline Monorail (Cyan Line)”. Two major nodes are clearly marked in crisp white text: “Tech District Central” near a high-density grid pattern, and “Grand Shrine Plaza” near a park symbol. At the bottom-center of the glass screen, a pulsing green button reads, “Tap here to plan your journey”. Near the lower-right, a detailed panel titled “Transfer & Accessibility Index” lists station features, including the line, “Wheelchair accessible transfer via elevators”. In the top-right corner, a small digital clock reads “14:32 | Oct 12, 2026”. Soft, colorful reflections of city lights and passing commuters blur across the glass surface of the display.

Test 09. A medium close-up shot of a modern living room television screen at night, the room warmly lit by a single floor lamp, casting cozy shadows. The TV screen is filled with a dynamic election night broadcast featuring a split-screen layout. On the left side of the split screen, a title block reads “GEORGIA RUNOFF UPDATE” in elegant, white serif letters. On the right side, a glowing red box in the upper corner displays “BREAKING RESULTS” in bold, white, uppercase sans-serif. In the middle of the screen, the razor-thin margin is displayed as “Marcus Vance: 50.1%” next to a blue bar and “Elena Rostova: 49.9%” next to a red bar. Directly beneath these statistics, a bright orange warning graphic reads “TOO CLOSE TO CALL” in bold black letters. At the bottom of the broadcast, a dark blue lower-third banner presents the headline “Voters line up in record numbers across Atlanta suburbs” in crisp, clean white letters, while a fast-moving ticker below displays “FULTON COUNTY REPORTS 100% OF PRECINCTS - RECOUNT EXPECTED” in a small, bright yellow digital font.

Test 10. A dramatic medium shot of a historical peace treaty document laid out on a crimson velvet tablecloth in a grand hall, illuminated by warm, flickering candle flame. The treaty is written on yellowed vellum in elaborate, ornate copperplate calligraphy. The main header at the top of the page reads “THE TREATY OF WESTMINSTER-ON-THAMES (1814)”. Directly below, the first section header states “ARTICLE IX: CESSATION OF HOSTILITIES AND EXCHANGE OF PRISONERS”. The core treaty clause reads “All offensive military actions by land or sea shall cease immediately upon the ratification of this instrument”. The secondary appendix page reads “ANNEX C: DEMARCATION OF MARITIME BOUNDARIES AND FISHING RIGHTS”. In the lower-right quadrant, the signatory block is decorated with a gold-embossed seal, reading “In representation of His Majesty: Lord Charles Sterling, Earl of Danby” and “In representation of the Regency: Count Henri de Valois”, with elaborate red wax seals and dangling silk ribbons finishing the presentation.