DexOPD: Dual-Extrapolation On-Policy Distillation for Diffusion Models

Yunchao Ke[1,2,*,‡], **Yanzhe Hu**[1,*], **Yi Yang**[1], **Jiachun Jin**[1], **Zhihong Liu**[1], **Zhijie Deng**[1,†]

[1]Shanghai Jiao Tong University, [2]Nankai University

*Equal contribution, ‡Work done during an internship at Shanghai Jiao Tong University, †Corresponding author

Flow-based diffusion models have become a mainstream paradigm for text-to-image generation, where task-specific post-training produces specialist models with complementary strengths. Multi-teacher on-policy distillation (OPD) offers an efficient way to consolidate these specialists into a single unified student. Recent methods strengthen supervision through teacher extrapolation, constructing stronger proxy teachers that allow distilled students to surpass the original specialists. However, their extrapolation directions remain defined solely by frozen teacher and reference models, leaving the student's own training progress unexploited. To address this limitation, we introduce **DexOPD**, a dual-extrapolation on-policy distillation method for diffusion models. DexOPD retains extrapolated teacher supervision while leveraging a lagging historical student as the reference for student self-extrapolation. Sampling across multiple self-extrapolation gains from shared initial noise broadens exploration along the student's evolving trajectory, and task rewards convert pairwise differences in candidate quality into a normalized auxiliary regression direction. Empirically, we evaluate DexOPD on GenEval, OCR, and DrawBench. The resulting unified SD3.5-Medium student achieves a GenEval reward of **0.9736**, an OCR score of **0.9515**, and a DrawBench Aesthetic score of **6.4875**, surpassing the corresponding specialist teachers as well as all evaluated OPD and multi-task reinforcement learning baselines across these three benchmarks while preserving standard single-model inference.

Code: https://github.com/SJTU-DENG-Lab/DexOPD

Correspondence: Zhijie Deng[†]: zhijied@sjtu.edu.cn

Date: November 24, 2026

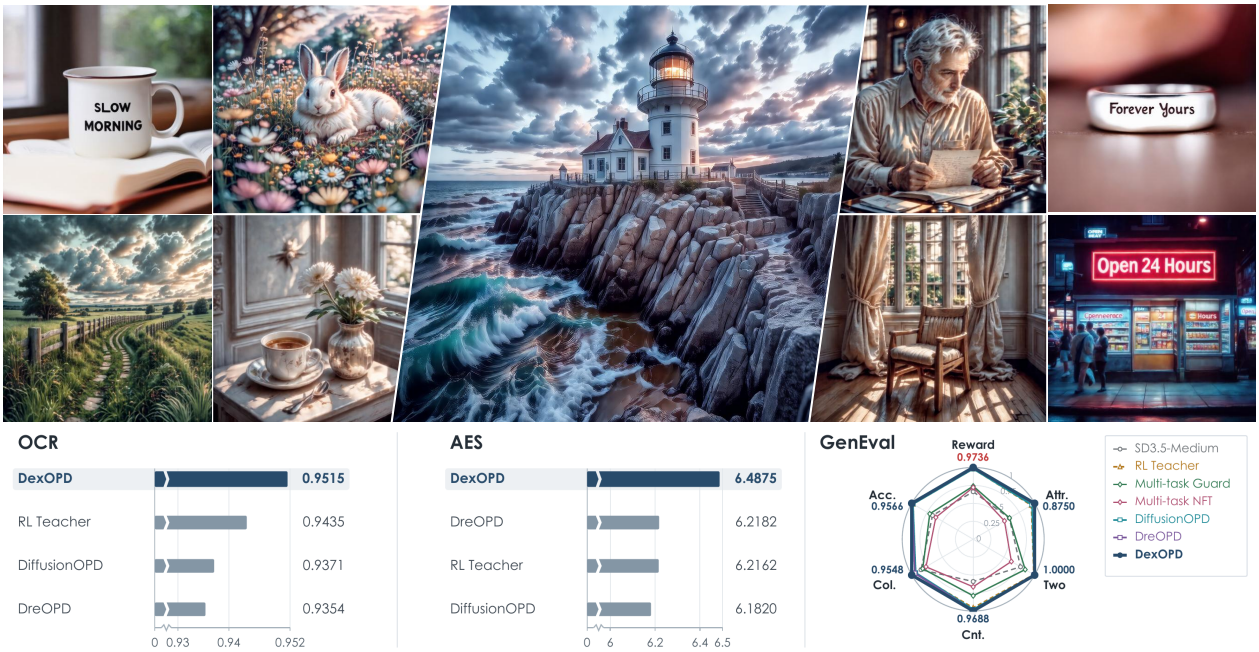


Figure 1 Qualitative highlights and multi-task benchmark performance. Top: representative images generated by our unified DexOPD student across compositional, text-rendering, and aesthetic prompts. Bottom: OCR and DrawBench Aesthetic (AES) scores are shown as horizontal bar charts; the radar chart reports six GenEval metrics normalized by each metric's maximum across the compared models, with raw DexOPD scores annotated.

1 Introduction

Flow-based diffusion models have become a leading paradigm for text-to-image generation (Lipman et al., 2023; Esser et al., 2024). Recent advances in post-training have substantially improved their performance on compositional generation, text rendering, and visual preference alignment. Optimizing separate models via task-specific reinforcement learning (RL) yields specialist experts with complementary strengths (Liu et al., 2025; Wang et al., 2026; Zheng et al., 2026). Multi-teacher on-policy distillation (OPD) provides an efficient mechanism to consolidate these RL-trained teachers into a single unified student: along its own denoising trajectories, the student learns from the corresponding task specialist at every visited state, aiming to match each expert on its respective task (Li et al., 2026a).

To push student performance beyond mere imitation, recent approaches extrapolate beyond each task teacher along the velocity difference between the teacher and a weaker reference model, constructing a stronger proxy teacher (Lin et al., 2026; Wei et al., 2026). Regressing toward these extrapolated velocity targets enables the student to outperform the original specialists on their respective benchmarks. Yet the extrapolation direction remains anchored by frozen teacher and reference models throughout training. As distillation progresses, the unified student integrates complementary skills across multiple tasks, but these newly acquired capabilities are not explicitly harnessed to generate further supervision. As illustrated in Figure 2, existing OPD methods surpass the original GenEval specialist yet plateau below the extrapolated proxy-teacher reference during late-stage training. This observation raises a central question: *can the student's own training progress provide additional, adaptive supervision to break past the ceiling of frozen teacher extrapolation?*

Our key insight is to *retain extrapolated teacher supervision while introducing historical student self-extrapolation to construct complementary supervision that evolves alongside training.* Specifically, the velocity difference between the current student and a lagging historical copy of itself captures the direction of recent optimization progress. Extrapolating along this self-difference from shared initial noise generates a diverse set of candidate endpoints that probe potential improvements grounded in the student's current competence. Because recent parameter updates do not uniformly improve every prompt, we evaluate these same-noise candidates with task-specific reward scorers and aggregate their pairwise quality differences into an auxiliary regression target on the native student trajectory.

Building on this insight, we propose **DexOPD**, a dual-extrapolation on-policy distillation framework for flow-based diffusion models. DexOPD jointly optimizes a unified student under two complementary branches: extrapolated teacher supervision, which anchors multi-task expertise from initialization onward, and reward-guided auxiliary supervision derived from historical student self-extrapolation, which activates naturally as the student evolves. After training, all teachers, historical references, and reward models are discarded, incurring zero architectural or computational overhead at inference time. Across GenEval, OCR, and DrawBench (Figure 1), our unified SD3.5-Medium student consistently outperforms both the individual specialist teachers and all evaluated OPD and multi-task RL baselines within a fixed 1,000-update budget.

In summary, our main contributions are threefold:

- **Historical Student Self-Extrapolation for OPD.** We introduce historical student self-extrapolation into multi-teacher on-policy distillation for flow-based diffusion models, leveraging the velocity difference between the current student and its lagging EMA copy to generate evolving, same-noise candidates beyond fixed teacher–reference directions.
- **The DexOPD Framework.** We propose **DexOPD**, a dual-extrapolation distillation algorithm that pairs extrapolated teacher targets with reward-guided covariance regression from self-extrapolated student candidates, jointly optimizing a unified student with zero inference-time overhead.
- **State-of-the-Art Multi-Task Consolidation.** Under a 1,000-update budget, our single unified student achieves **0.9736** GenEval reward, **0.9515** OCR, and **6.4875** DrawBench Aesthetic, surpassing the individual specialist teachers as well as all evaluated OPD and multi-task RL baselines across these three objectives.

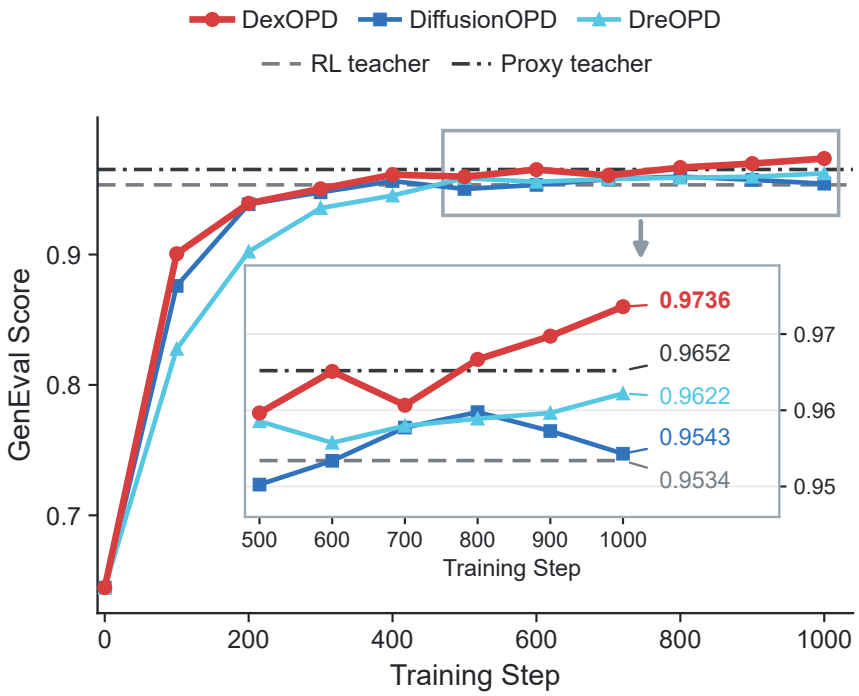


Figure 2 The student need not trail the proxy teacher. GenEval learning curves for three multi-teacher OPD methods over 1,000 optimizer updates. The inset magnifies the highlighted late-training interval (updates 500–1,000), comparing final scores against the native RL teacher and extrapolated proxy-teacher references.

2 Preliminaries

2.1 Flow Matching

Let c denote a text prompt, x_{clean} a clean image latent, and $\epsilon \sim \mathcal{N}(0, I)$ Gaussian noise. Flow matching learns a time-dependent velocity field transporting samples along a probability path between noise and data (Lipman et al., 2023; Esser et al., 2024). Under the linear interpolation path,

$$x_\sigma = (1-\sigma)x_{\text{clean}} + \sigma\epsilon, \qquad u_\sigma = \epsilon - x_{\text{clean}}, \tag{1}$$

and the velocity network v_θ minimizes the squared error $\mathbb{E}[\|v_\theta(x_\sigma, \sigma, c) - u_\sigma\|_2^2]$. Sampling integrates this learned velocity field from noise ($\sigma_0 = 1$) to data ($\sigma_N = 0$). Throughout this work, we adopt the deterministic Euler transition operator:

$$\mathcal{S}_j(x, v) = x + \Delta\sigma_j v, \qquad \Delta\sigma_j = \sigma_{j+1} - \sigma_j < 0, \tag{2}$$

where $j \in \{0, \ldots, N-1\}$ indexes the denoising step. The same transition operator $\mathcal{S}_j$ is used consistently for trajectory rollouts, student predictions, and regression target construction.

2.2 Multi-Teacher OPD with Extrapolation

For each task q, let T_q denote a frozen specialist teacher and B the frozen base model. In multi-teacher on-policy distillation, the student rolls out its own denoising trajectory for a prompt c from task q, and teachers are queried at the visited student states $s_j = (x_j, \sigma_j, c)$. Teacher extrapolation constructs an enhanced proxy velocity target by extrapolating beyond T_q along the task-adaptation direction $v_{T_q} - v_B$:

$$v_{\text{proxy},q}(s_j) = v_{T_q}(s_j) + \lambda\big(v_{T_q}(s_j) - v_B(s_j)\big), \tag{3}$$

where $\lambda \geq 0$ controls the extrapolation strength and $\lambda = 0$ recovers direct teacher supervision. Both the teacher and base predictions are evaluated with the same task-specific classifier-free guidance (CFG) scale, while the student advances its rollout exclusively under its own velocity field.

Following the transition-mean regression formulation of DiffusionOPD (Li et al., 2026a), we define the student transition prediction $\mu_{\theta,j} = \mathcal{S}_j(x_j, v_\theta(s_j))$ and the corresponding extrapolated teacher target:

$$\begin{aligned} \mu_{T,j} &= \mathrm{sg}\big[\mathcal{S}_j\big(x_j, v_{\mathrm{proxy},q}(s_j)\big)\big], \\ \mathcal{L}_{\mathrm{OPD}} &= \mathbb{E}_{q,c,\epsilon,j}\left[\tfrac{1}{2}\,\mathrm{MSE}(\mu_{\theta,j}, \mu_{T,j})\right], \end{aligned} \tag{4}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, j ranges over the supervised denoising steps $\mathcal{J}$, and MSE averages over latent dimensions. Because sampled states and teacher targets are cached and detached prior to optimization, gradients flow exclusively through the current student predictions $\mu_{\theta,j}$.

2.3 Reward-Guided Direction Aggregation

Given K candidate endpoint latents $\{z^{(k)}\}_{k=0}^{K-1}$ generated for the same prompt c, a VAE decoder D and task-specific reward model R_q yield scalar scores $r_k = R_q(D(z^{(k)}), c)$. To translate these black-box reward signals into a dense latent update direction, we adopt the endpoint–reward covariance aggregation of DiffusionART (Wan et al., 2026):

$$\begin{aligned} \bar{z} &= \tfrac{1}{K}\sum_{k=0}^{K-1} z^{(k)}, \qquad \bar{r} = \tfrac{1}{K}\sum_{k=0}^{K-1} r_k, \\ C &= \tfrac{1}{K}\sum_{k=0}^{K-1} (z^{(k)} - \bar{z})(r_k - \bar{r}). \end{aligned} \tag{5}$$

Let $s_C = \sqrt{\mathrm{mean}(C^2)}$ denote the root-mean-square (RMS) magnitude of C across latent elements, and define the activity indicator $a = \mathbf{1}[\max_k r_k \neq \min_k r_k]\,\mathbf{1}[s_C > \varepsilon_C]$. The normalized auxiliary direction is then given by

$$G = \begin{cases} C/s_C, & a = 1, \\ 0, & a = 0. \end{cases} \tag{6}$$

Notably, C depends solely on relative differences across candidates and admits the equivalent pairwise form:

$$C = \frac{1}{K^2} \sum_{0 \le k < \ell < K} (z^{(k)} - z^{(\ell)})(r_k - r_\ell). \tag{7}$$

Consequently, both higher-reward and lower-reward candidates contribute constructively through their pairwise contrasts, whereas groups with identical rewards produce $G = 0$. Whenever active ($a = 1$), G has unit RMS magnitude and requires only black-box scorer evaluations. We use Equations (5) and (6) to translate reward contrasts among historically self-extrapolated student candidates into a normalized update direction that complements multi-teacher OPD.

3 Method

Teacher extrapolation improves distillation targets by extrapolating beyond each specialist along a frozen teacher–base velocity difference. As training progresses, however, the unified student acquires multi-task capabilities that neither any individual teacher nor the base model fully captures. DexOPD leverages this evolving student velocity field as a complementary source of supervision alongside extrapolated teacher guidance. At the start of each collection round r, let θ_r denote the current student parameters and H_r its historical reference. Native rollouts from θ_r yield on-policy states supervised by extrapolated task teachers, while self-extrapolation against H_r produces same-noise candidates whose task rewards induce an auxiliary corrective direction. Figure 3 illustrates the complete dual-extrapolation framework.

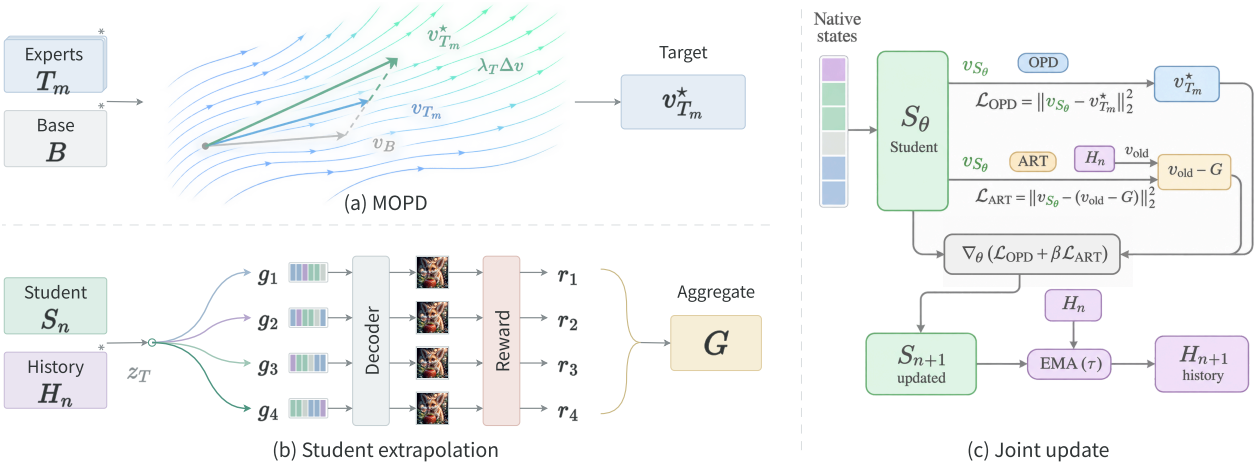


Figure 3 Overview of DexOPD. (a) Extrapolated task teachers supervise native student states via $\mathcal{L}_{\text{OPD}}$. (b) Historical self-extrapolation against H_r generates same-noise candidates whose endpoint–reward covariance produces the unit-RMS direction G. (c) Teacher and auxiliary targets jointly update the student before the historical EMA reference is refreshed. In (a), m indexes tasks; student and history subscripts in (b,c) index training rounds (denoted q and r in the text).

3.1 Candidate Generation via Historical Self-Extrapolation

We maintain the historical reference H_r as an exponential moving average (EMA) of the trainable student parameters, initialized at $H_0 = \theta_0$ and updated after each training round r:

$$H_{r+1} = \tau H_r + (1-\tau)\theta_{r+1}, \qquad 0 < \tau < 1. \tag{8}$$

This lagging reference anchors the student's recent optimization trajectory, summarizing both previously distilled teacher knowledge and prior reward-guided updates. Motivated by historical self-extrapolation (Li et al., 2026b), we extrapolate the velocity difference between the current student and its historical reference at any state (x, σ, c):

$$v_r^{(g)}(x,\sigma,c) = v_{\theta_r}(x,\sigma,c) + g\big(v_{\theta_r}(x,\sigma,c) - v_{H_r}(x,\sigma,c)\big). \tag{9}$$

Given a schedule of extrapolation gains $\mathcal{G} = \{0, g_1, \ldots, g_{K-1}\}$, we sample K candidate trajectories starting from the same initial noise $x_0^{(g)} = \epsilon$ and text prompt c:

$$\begin{aligned} x_0^{(g)} &= \epsilon, \\ x_{j+1}^{(g)} &= \mathcal{S}_j\left(x_j^{(g)},\, v_r^{(g)}(x_j^{(g)}, \sigma_j, c)\right), \\ z^{(g)} &= x_N^{(g)}. \end{aligned} \tag{10}$$

The $g = 0$ candidate directly reuses the native student endpoint already computed for OPD, requiring no extra sampling. Positive gains ($g > 0$) probe varying extrapolation magnitudes under identical initial conditions.

Structurally, Equation (9) mirrors teacher extrapolation in Equation (3)—both take the form $v_{\text{current}} + \gamma(v_{\text{current}} - v_{\text{ref}})$—but replaces the frozen base model B with the evolving historical student H_r. As a result, the reference itself tracks the student's recent progress. Because candidates are generated by extrapolating along the student's actual optimization trajectory rather than random perturbations, they explore directions the model is actively moving toward; reward-ranked contrasts among them are therefore well-suited to isolate genuine improvements grounded in the student's current competence.

3.2 Reward-Guided Auxiliary Target Construction

Each candidate endpoint $z^{(g)}$ is decoded into image space and scored by the corresponding task reward model R_q, yielding the normalized covariance direction G via Equations (5) and (6). Let $v_{\text{old},j} = v_{\theta_r}(s_j)$ denote the

cached velocity of the native trajectory at supervised state s_j. We construct the one-step auxiliary regression target as

$$\mu_{\text{aux},j} = \text{sg}[\mathcal{S}_j(x_j,\, v_{\text{old},j} - G)]\,. \tag{11}$$

Under our noise-to-data integration convention ($\Delta\sigma_j < 0$), subtracting G from the velocity shifts the target state along $+G$ with RMS displacement $|\Delta\sigma_j|$; the complete algebraic derivation and blended target form are provided in Appendix B.

Because G is formed from pairwise reward contrasts (Equation (7)), candidate pairs that differ in quality determine both the sign and orientation of the update: lower-reward candidates suppress undesirable velocity shifts while higher-reward candidates reinforce beneficial ones, all projected cleanly onto the native student trajectory states.

3.3 Joint Dual-Extrapolation Optimization

The teacher and self-extrapolation branches exhibit naturally complementary dynamics across training. At initialization ($H_0 = \theta_0$), all candidate velocities coincide with the native trajectory ($G = 0$), allowing extrapolated teacher supervision to bootstrap multi-task competence cleanly without noisy early exploration. As θ_r diverges from H_r, reward differences across same-noise candidates activate the auxiliary branch $\mathcal{L}_{\text{ART}}$, supplying adaptive, self-discovered supervision even after the student matches or surpasses individual specialist teachers.

We weight the auxiliary transition loss by the noise scale σ_j and combine it with extrapolated teacher distillation:

$$\begin{aligned} \mathcal{L}_{\text{ART}} &= \mathbb{E}_{q,c,\epsilon,j}\left[\sigma_j a\ \text{MSE}(\mu_{\theta,j}, \mu_{\text{aux},j})\right], \\ \mathcal{L} &= \mathcal{L}_{\text{OPD}} + \beta\,\mathcal{L}_{\text{ART}}, \end{aligned} \tag{12}$$

where $\beta > 0$ balances teacher distillation against student self-evolution feedback, and the expectation averages over all samples in the batch (including inactive groups with $a = 0$). All native rollouts, candidate endpoints, scalar rewards, and regression targets are pre-computed and detached prior to gradient updates; gradients backpropagate solely through the student predictions $\mu_{\theta,j}$.

3.4 Round-Based Training and Zero-Overhead Inference

As summarized in Algorithm 1, each training round freezes θ_r and H_r to collect task-balanced native trajectories and self-extrapolated candidate groups, pre-computing both $\mu_{T,j}$ and $\mu_{\text{aux},j}$ before performing optimizer updates. Once the round's updates conclude, the historical reference is refreshed via Equation (8). After training finishes, all specialist teachers, historical EMA references, and reward scorers are discarded; at test time, the unified student executes standard single-trajectory sampling with zero architectural or computational overhead. Complete scheduling, caching, and hyperparameter details are provided in Appendix A.

Algorithm 1 DexOPD: Dual-Extrapolation On-Policy Distillation

Require: Frozen specialist teachers $\{T_q\}$, base model B, task prompts, reward scorers $\{R_q\}$, teacher strength λ, auxiliary weight β, candidate gains $\mathcal{G}$, EMA decay τ

Initialize student adapter θ and historical reference $H \leftarrow \theta$
while update budget is not exhausted **do**
  Freeze sampling parameters $\theta_r \leftarrow \theta$, $H_r \leftarrow H$; clear round cache
  for each task-balanced sample (q, c, ϵ) **do**
    Roll out native student trajectory with θ_r; cache states $\{s_j\}$ and velocities $\{v_{\text{old},j}\}$
    Construct extrapolated teacher targets $\{\mu_{T,j}\}$ via Equations (3) and (4)
    Reuse $g = 0$ endpoint; sample positive-gain candidates $g \in \mathcal{G} \setminus \{0\}$ via Equations (9) and (10)
    Score decoded candidates with R_q; compute G (Equations (5) and (6)) and targets $\{\mu_{\text{aux},j}\}$ (Equation (11))
  end for
  Shuffle cached records into mini-batches; update θ by minimizing Equation (12)
  Refresh historical reference: $H \leftarrow \tau H + (1-\tau)\theta$
end while
return Unified student parameters θ

4 Experiments

We conduct comprehensive experiments to evaluate DexOPD on multi-teacher consolidation for text-to-image generation. We first compare our unified student against single-task specialist teachers, multi-task reinforcement learning methods, and recent on-policy distillation baselines across quantitative metrics, learning dynamics, and qualitative outputs. We then isolate the contributions of dual extrapolation, historical referencing, and auxiliary loss weighting through controlled ablations, followed by an analysis of inference-time teacher extrapolation.

4.1 Experimental Setup

Following DiffusionOPD (Li et al., 2026a), we adopt SD3.5-Medium (Esser et al., 2024) as the base backbone and train at 512×512 resolution using the GenEval, OCR, and Pick-a-Pic (Kirstain et al., 2023) prompt splits. We evaluate compositional alignment on GenEval (Ghosh et al., 2023), text rendering on OCR, and visual quality on DrawBench via Aesthetic, PickScore (Kirstain et al., 2023), ImageReward (Xu et al., 2023), CLIPScore (Hessel et al., 2021), and HPSv2.1 (Wu et al., 2023). Full hyperparameter, CFG, prompt-split, and baseline implementation details are reported in Appendix A.

Optimization. We train a rank-32 LoRA adapter on the student while keeping the diffusion backbone, text encoders, VAE, and task teachers frozen. Each collection round gathers 144 task-balanced samples (48 per task) and performs two optimizer updates with mini-batches of 72, ensuring equal teacher representation within every round. We run 10 deterministic Euler sampling steps and supervise the first nine states, using the final transition to produce the endpoint for reward scoring. All candidates for a given prompt share identical initial noise, and the native OPD rollout is reused directly as the $g = 0$ candidate. Both teacher and auxiliary regression targets are cached prior to the two updates so that supervision remains consistent throughout each round. Our default configuration uses teacher extrapolation $\lambda = 1.2$, auxiliary loss weight $\beta = 1.8$, candidate gains $\mathcal{G} = \{0, 0.05, 0.25, 0.5\}$, and history EMA decay $\tau = 0.995$ per round. All methods are evaluated after an identical budget of 1,000 optimizer updates.

Single-Task Teachers. We use the three frozen task specialists released by DiffusionOPD (Li et al., 2026a): a GenEval teacher trained with DiffusionNFT (Zheng et al., 2026), and OCR and Aesthetic teachers trained with GRPO-Guard (Wang et al., 2026). The Aesthetic specialist optimizes PickScore, CLIPScore, and HPSv2.1.

Baselines. We benchmark DexOPD against the base SD3.5-Medium model, the three single-task specialist teachers, multi-task RL baselines—GRPO-Guard (Wang et al., 2026), DiffusionNFT (Zheng et al., 2026), and DiffusionART (Wan et al., 2026)—and multi-teacher OPD baselines: DiffusionOPD (Li et al., 2026a), DreOPD (Lin et al., 2026), and our paper-based reproduction of STEP-OPD (Wei et al., 2026).

4.2 Multi-Task Generation Results

Quantitative Comparison. As shown in Table 1, DexOPD surpasses all three original specialist teachers as well as every evaluated OPD and multi-task RL baseline across all three primary objectives, and further exceeds the extrapolated proxy-teacher references on GenEval (Figure 2) and DrawBench Aesthetic while avoiding the severe cross-task degradation of standalone teacher extrapolation (Table 4). Under the same 1,000-update budget, multi-task RL baselines attempt joint optimization directly but trail DexOPD by a wide margin. Concretely, our unified student reaches **0.9736** GenEval reward, **0.9515** OCR, and **6.4875** Aesthetic, compared with 0.9622, 0.9354, and 6.2182 for DreOPD. On GenEval, DexOPD achieves an overall accuracy of 0.9566, delivering clear gains over DreOPD on counting (0.9688 vs. 0.9627) and color attribution (0.9548 vs. 0.9069), while DreOPD maintains a higher position score. Whereas the Aesthetic specialist is trained on PickScore, CLIPScore, and HPSv2.1, our auxiliary branch directly optimizes the DrawBench

Table 1 Multi-task generation performance after 1,000 updates. GenEval sub-categories and Acc. report classification accuracies; Reward denotes the dense GenEval reward score. AES, Pick, IR, CLIP, and HPS correspond to Aesthetic, PickScore, ImageReward, CLIPScore, and HPSv2.1. Higher is better across all columns. Bold and underline denote the best and second-best results, respectively. *: paper-based reproduction.

Model	GenEval								OCR	DrawBench				
	Sing.	Two	Cnt.	Col.	Pos.	Attr.	Acc.	Reward		AES	Pick	IR	CLIP	HPS
(a) Base model														
SD3.5-Medium	0.9844	0.7702	0.5714	0.8112	0.2200	0.5149	0.6336	0.6446	0.5486	5.3899	0.8614	0.8383	0.2895	0.2924
(b) Single-task specialist teachers														
GenEval	**1.0000**	0.9848	0.9250	0.9069	**0.9750**	0.8275	0.9349	0.9534	0.6515	5.2635	0.8513	0.9277	0.2869	0.2673
OCR	0.9719	0.7172	0.6000	0.7819	0.2225	0.4725	0.6144	0.6617	0.9435	5.2605	0.8580	0.8972	0.2909	0.2732
Aesthetic	0.9719	0.8864	0.6063	0.8670	0.2950	0.6575	0.7066	0.4791	0.5881	6.2162	0.9233	1.5016	0.2960	0.3452
(c) Multi-task reinforcement learning														
GRPO-Guard	0.9875	0.8460	0.7640	0.7846	0.2475	0.5174	0.6764	0.7202	0.7210	5.6812	0.8770	1.1531	0.2904	0.3091
DiffusionNFT	0.8844	0.6212	0.6429	0.7340	0.2450	0.4453	0.5817	0.7029	0.9434	5.3674	0.8412	0.6314	0.2787	0.2642
DiffusionART	0.8406	0.5707	0.5497	0.5665	0.1900	0.2811	0.4847	0.4427	0.9033	5.6673	0.8292	0.7193	0.2710	0.2653
(d) Multi-teacher on-policy distillation														
DiffusionOPD	**1.0000**	0.9924	0.9472	0.8936	0.9650	0.8383	0.9373	0.9543	0.9371	6.1820	0.9237	1.5110	**0.3006**	0.3547
DreOPD	**1.0000**	0.9899	0.9627	0.9069	0.9725	**0.8756**	0.9495	0.9622	0.9354	6.2182	0.9237	**1.5259**	0.2986	0.3558
STEP-OPD*	**1.0000**	0.9949	0.9379	0.8989	**0.9750**	0.8657	0.9440	0.9562	0.9378	6.2813	**0.9248**	1.5207	0.2985	**0.3566**
DexOPD (ours)	**1.0000**	**1.0000**	**0.9688**	**0.9548**	0.9525	0.8750	**0.9566**	**0.9736**	**0.9515**	**6.4875**	0.9190	1.5127	0.2914	0.3484

Aesthetic scorer, lifting Aesthetic substantially (6.4875 vs. 6.2813 for STEP-OPD*) while keeping PickScore (0.9190) and HPSv2.1 (0.3484) closely aligned with the strongest OPD baselines.

The specialist teachers also underscore the severity of multi-task trade-offs: the GenEval teacher reaches 0.9534 on GenEval reward yet drops to 0.6515 on OCR, whereas the OCR teacher achieves 0.9435 on OCR but only 0.6617 on GenEval reward. Our single unified student exceeds both specialist scores simultaneously, while raising DrawBench Aesthetic well above the Aesthetic teacher's 6.2162. Furthermore, as detailed in Appendix F, a small-batch variant of DexOPD (batch size 24) matches the wall-clock time per update of standard DiffusionOPD while still outperforming all OPD baselines across all three target metrics.

Learning Dynamics. On GenEval, DexOPD separates from competing OPD methods early and climbs past both the native and proxy-teacher references, whereas DiffusionOPD and DreOPD plateau below the proxy teacher (Figure 2). Across all three tasks simultaneously, multi-task DiffusionNFT and DiffusionART initially make progress on GenEval before degrading mid-training, whereas DexOPD improves steadily and reliably throughout the 1,000-update horizon (Figure 4). This contrast underscores that the fixed-budget gains reflect superior optimization stability in addition to stronger final endpoints.

Qualitative Comparison. Figure 5 compares generated images across six prompts under shared initial noise. Relative to the individual specialist teachers and OPD baselines, DexOPD more reliably satisfies fine-grained object composition, renders intricate poster and console text legibly, and preserves rich visual textures and comic details within a single unified model. Additional generation examples across 77 GenEval and 88 OCR prompts are presented in Appendix G.

4.3 Ablation Studies and Analysis

Component Ablations. Table 2 isolates each component of DexOPD under identical optimization budgets and schedules. Removing teacher extrapolation ($\lambda = 0$) lowers all three primary target scores substantially (by 0.0221 on GenEval, 0.0158 on OCR, and 0.1945 on Aesthetic). Holding teacher extrapolation fixed at $\lambda = 1.2$ and removing the student self-extrapolation branch ($\beta = 0$) still degrades performance across all three tasks,

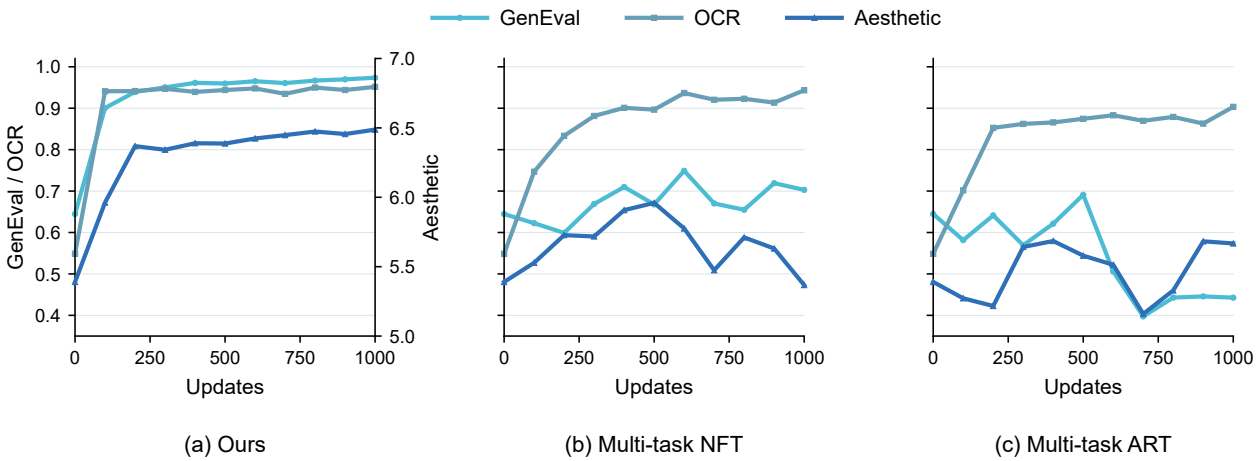


Figure 4 Three-task learning dynamics over 1,000 optimizer updates. Curves compare (a) DexOPD, (b) multi-task DiffusionNFT, and (c) multi-task DiffusionART. The left vertical axis measures GenEval and OCR scores; the right vertical axis in (a) measures DrawBench Aesthetic.

reducing GenEval reward by 0.0067, OCR by 0.0107, and Aesthetic by 0.0465. Finally, replacing the evolving historical reference H_r with the frozen base model B degrades performance across tasks—most prominently dropping Aesthetic from 6.4875 to 6.2548 and GenEval from 0.9736 to 0.9633—confirming that anchoring self-extrapolation to an evolving EMA student history is essential.

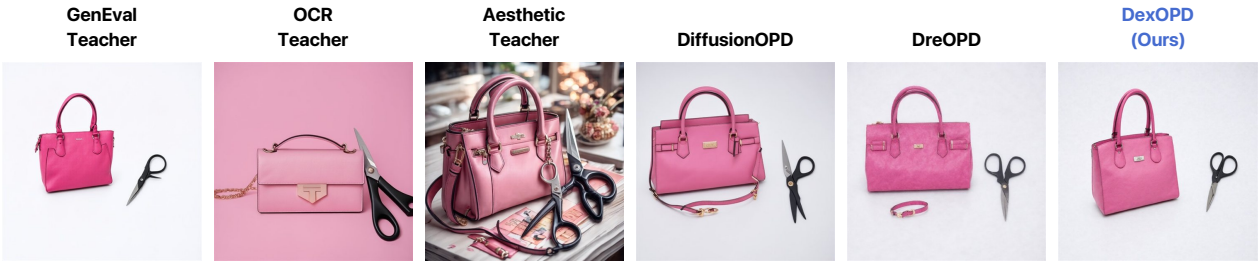
Table 2 Component ablation at 1,000 updates. Each variant modifies a single design element of DexOPD under identical optimization schedules. Bold marks the best result per column, and the highlighted row denotes our default configuration.

Configuration	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
DexOPD (default)	**0.9736**	**0.9515**	**6.4875**	0.9190	**1.5127**	0.2914	0.3484
No teacher extrapolation ($\lambda = 0$)	0.9515	0.9357	6.2930	**0.9215**	1.4903	**0.2960**	0.3366
No student auxiliary ($\beta = 0$)	0.9669	0.9408	6.4410	0.9185	1.4957	0.2944	**0.3527**
Frozen-base reference ($H_r = B$)	0.9633	0.9512	6.2548	0.8844	1.4034	0.2843	0.3254

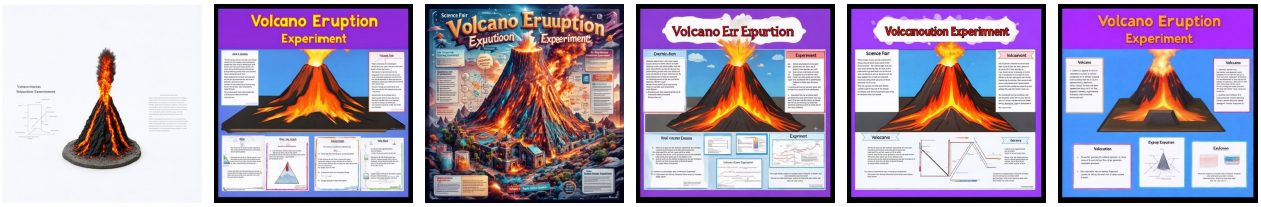
Candidate Feedback and Activity Analysis. No single fixed self-extrapolation gain g monotonically improves mean reward across all three tasks; instead, covariance aggregation (Equation (7)) extracts useful supervision whenever candidates exhibit reward variation. Across the full default run (24,000 four-candidate groups per task), the auxiliary branch is active in 9.50% of GenEval groups, 26.44% of OCR groups, and 99.80% of Aesthetic groups (Appendix C reports full per-gain win/tie/loss statistics and training diagnostics). Crucially, covariance aggregation harvests productive supervision directly from reward diversity across gains: at $g = 0.5$, Aesthetic candidates win 39.87% and lose 59.93% of same-noise comparisons, yet symmetric pairwise contrasts convert both higher- and lower-reward outcomes into a consistent corrective direction.

Auxiliary-Weight Sensitivity. Table 3 summarizes how the auxiliary loss weight β modulates cross-task trade-offs (with the complete 11-point sweep reported in Appendix D). Among the displayed settings, $\beta = 0.4$ attains the highest GenEval reward (0.9743), $\beta = 2.0$ yields the highest OCR score (0.9520), and our default $\beta = 1.8$ achieves the highest Aesthetic score (6.4875) and ImageReward (1.5127) while remaining within 0.0007 of the best GenEval score and 0.0005 of the best OCR score. Every positive weight improves Aesthetic and ImageReward over $\beta = 0$, whereas GenEval and OCR vary non-monotonically; $\beta = 1.8$ thus serves as a well-balanced operating point across all three primary objectives.

Teacher Inference Extrapolation. To contextualize the limits of frozen teacher extrapolation, Figure 6 plots on-task performance as a function of inference extrapolation strength λ, and Table 4 reports multi-metric



a photo of a pink handbag and a black scissors



A detailed science fair poster titled "Volcano Eruption Experiment", showcasing a colorful, explosive model of a volcano with lava spilling down its sides, surrounded by diagrams and facts about volcanic eruptions, all set against a vibrant, educational background.



A courtroom scene with a judge's gavel resting on a wooden plaque that reads "Order in the Court", set against the backdrop of a quiet, solemn courtroom.



A futuristic sci-fi spaceship console with sleek, metallic surfaces and glowing blue panels. The console's central display blinks "Hyperdrive Ready" in bright, neon letters, surrounded by an array of blinking lights and control buttons.



a photo of a toilet and a computer mouse



A magnifying glass over a page of a 1950s batman comic.

Figure 5 Qualitative comparison on six representative prompts. From left to right, each row shows outputs from the GenEval, OCR, and Aesthetic specialist teachers, DiffusionOPD, DreOPD, and DexOPD generated from shared initial noise.

Table 3 Auxiliary-weight ablation at 1,000 updates. Selected values of β with $\lambda = 1.2$ fixed; the complete 11-point sweep is provided in Appendix D. Bold and underline mark the best and second-best values per column.

β	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
0.0	0.9669	0.9408	6.4410	0.9185	1.4957	**0.2944**	**0.3527**
0.4	**0.9743**	0.9414	6.4533	0.9190	1.5036	0.2911	0.3478
0.8	0.9647	0.9452	6.4689	0.9192	1.5045	0.2919	0.3477
1.2	0.9717	0.9500	6.4718	<u>0.9198</u>	<u>1.5123</u>	0.2919	0.3481
1.6	0.9526	0.9437	6.4622	**0.9201**	1.5118	0.2919	0.3470
1.8	<u>0.9736</u>	<u>0.9515</u>	**6.4875**	0.9190	**1.5127**	0.2914	<u>0.3484</u>
2.0	0.9587	**0.9520**	<u>6.4728</u>	0.9187	1.5108	<u>0.2920</u>	0.3471

trade-offs across selected strengths (the complete sweep appears in Appendix E). At $\lambda = 0.8$, inference extrapolation boosts each specialist on its own objective: GenEval rises from 0.9534 to 0.9651, OCR from 0.9435 to 0.9616, and Aesthetic from 6.2162 to 6.4164. However, these on-task gains incur severe cross-task degradation—for example, the OCR teacher's GenEval score drops from 0.6617 to 0.5613—and overly aggressive extrapolation degrades even the target metric itself. At $\lambda = 2.0$, the Aesthetic teacher drops from its 6.4164 peak to 6.0806 (with GenEval falling to 0.2288), while the OCR teacher falls to 0.9487 OCR and negative ImageReward (-0.2367). Thus, naively increasing teacher extrapolation strength neither uniformly improves the target task nor preserves off-task quality. For student distillation, we fix $\lambda = 1.2$ across all tasks—a balanced operating point where each teacher remains close to its on-task peak (within 0.0009 on GenEval, 0.0057 on OCR, and 0.0117 on Aesthetic) while avoiding the collapse observed at larger strengths.

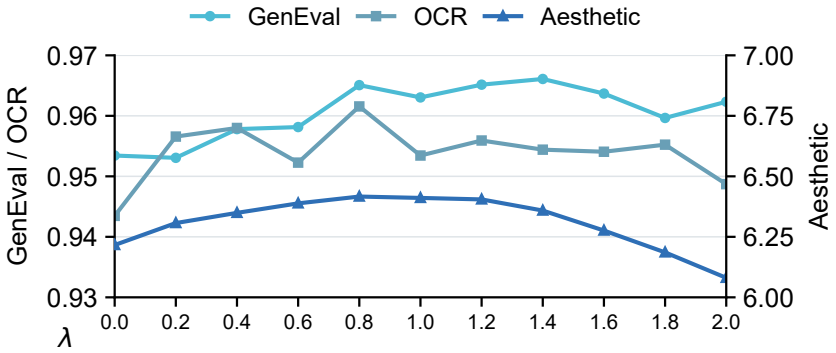


Figure 6 Inference-time teacher extrapolation. Each specialist RL teacher is evaluated on its own target metric as extrapolation strength λ varies from 0.0 to 2.0. On-task scores peak around $\lambda \in [0.8, 1.4]$ and degrade under excessive extrapolation.

5 Related Work

Multi-Teacher Distillation for Image Generation. DiffusionOPD decouples specialist acquisition from multi-task consolidation by supervising a unified student along its own on-policy trajectories (Li et al., 2026a). Flow-OPD and DanceOPD further explore on-policy alignment for flow-matching models and generative field integration (Fang et al., 2026; Zhou et al., 2026b). To move beyond direct teacher imitation, DreOPD constructs extrapolated supervision using a degraded reference model (Lin et al., 2026), while STEP-OPD combines output extrapolation with internal representation alignment (Wei et al., 2026). While these methods strengthen the targets derived from frozen teachers, DexOPD retains extrapolated teacher supervision and complements it with an adaptive signal drawn directly from the unified student's own evolving velocity field.

Historical Extrapolation and Self-Distillation. In language modeling, RISE extrapolates a policy against a lagging historical copy in parameter or output space and integrates the resulting self-teacher with reward learning (Li

Table 4 Multi-metric effects of teacher inference extrapolation. Selected extrapolation strengths λ for each specialist teacher. Bold marks the best displayed on-task score per teacher. Higher is better for all metrics; the complete sweep appears in Appendix E.

Teacher	λ	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
GenEval	0.0	0.9534	0.6515	5.2635	0.8513	0.9277	0.2869	0.2673
	0.4	0.9578	0.4302	5.2781	0.8439	0.5093	0.2739	0.2458
	0.8	0.9651	0.4559	5.2781	0.8420	0.4716	0.2717	0.2452
	1.2	**0.9652**	0.4714	5.2611	0.8401	0.4643	0.2710	0.2429
	1.6	0.9637	0.4733	5.2501	0.8375	0.4041	0.2672	0.2415
	2.0	0.9623	0.4872	5.2526	0.8358	0.3536	0.2645	0.2387
OCR	0.0	0.6617	0.9435	5.2605	0.8580	0.8972	0.2909	0.2732
	0.4	0.5914	0.9580	5.1923	0.8357	0.5848	0.2891	0.2456
	0.8	0.5613	**0.9616**	5.1091	0.8262	0.4202	0.2855	0.2370
	1.2	0.5145	0.9559	5.0281	0.8166	0.2917	0.2807	0.2299
	1.6	0.4695	0.9541	4.8835	0.8055	−0.0227	0.2711	0.2177
	2.0	0.4276	0.9487	4.7608	0.7959	−0.2367	0.2646	0.2095
Aesthetic	0.0	0.4791	0.5881	6.2162	0.9233	1.5016	0.2960	0.3452
	0.4	0.5011	0.6017	6.3491	0.9228	1.5071	0.2959	0.3444
	0.8	0.4499	0.5686	**6.4164**	0.9170	1.4932	0.2940	0.3440
	1.2	0.3903	0.5269	6.4047	0.9068	1.4589	0.2928	0.3398
	1.6	0.3331	0.4981	6.2760	0.8935	1.3907	0.2924	0.3307
	2.0	0.2288	0.4858	6.0806	0.8759	1.2604	0.2891	0.3160

et al., 2026b), whereas SDPO performs feedback-conditioned self-distillation (Hübotter et al., 2026). Within image generation, D-OPSD builds self-teachers via enriched image–text conditioning for step-distilled diffusion models (Jiang et al., 2026), and DiffusionOPSD investigates reward-guided self-distillation targets (Zhou et al., 2026a). Our approach shares the perspective that a learner can generate valuable supervision for itself, but specifically instantiates historical self-extrapolation for same-noise candidate generation in continuous flow-matching velocity fields, pairing image-endpoint reward contrasts with external multi-teacher supervision.

Diffusion Reinforcement Learning. Flow-GRPO, GRPO-Guard, and DiffusionNFT optimize image rewards directly via online policy gradients, regulated clipping, and forward-process reinforcement, respectively (Liu et al., 2025; Wang et al., 2026; Zheng et al., 2026). DiffusionART aggregates endpoint latents and scalar rewards into a normalized regression direction, enabling stable learning with black-box reward scorers (Wan et al., 2026). Building on this covariance aggregation interface, DexOPD introduces historical self-extrapolation to generate same-noise candidates from the evolving student and fuses their task-specific feedback with extrapolated teacher targets directly on native on-policy trajectories.

6 Conclusion

We introduced **DexOPD**, a dual-extrapolation on-policy distillation framework for flow-based diffusion models that pairs extrapolated teacher supervision with reward-guided feedback from historical student self-extrapolation. Within 1,000 optimizer updates, our unified SD3.5-Medium student achieves a GenEval reward of **0.9736**, an OCR score of **0.9515**, and a DrawBench Aesthetic score of **6.4875**, surpassing the corresponding specialist teachers as well as all evaluated OPD and multi-task RL baselines on these three core objectives while preserving standard single-model inference with zero test-time overhead. Controlled ablations confirm that adding the self-extrapolation auxiliary branch on top of fixed teacher extrapolation improves GenEval, OCR, and Aesthetic by 0.0067, 0.0107, and 0.0465, respectively, and that anchoring self-extrapolation to an evolving EMA student history consistently outperforms a frozen base reference.

Limitations and Future Work. Sampling multiple self-extrapolated candidates increases per-round rollout compute relative to single-trajectory OPD; reducing this overhead via adaptive candidate scheduling or fewer sampling steps—building on our compute-matched small-batch regime (Appendix F)—is a natural next step. In addition, our current evaluation focuses on SD3.5-Medium across three core text-to-image objectives and their associated reward models; extending dual-extrapolation distillation to larger diffusion backbones, multi-reward preference ensembles, and video or image-editing tasks remains an exciting direction for future work.

References

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven C. H. Hoi. D-OPSD: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. DiffusionOPD: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026a.

Yang Li, Semih Yavuz, and Shafiq Joty. RISE: Recursive improvement via self-extrapolating policy distillation. *arXiv preprint arXiv:2609.05295*, 2026b.

Mingfeng Lin, Chengfei Cai, Lin Xu, Yuxiang Wei, and Liang Han. DreOPD: Degraded-reference extrapolative on-policy distillation for flow-matching models. *arXiv preprint arXiv:2608.09233*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhenglin Wan, Jiayu Chen, Zhengfeiyang Zhang, Pengfei Zhou, Wangbo Zhao, Fangyuan Kong, Pengfei Wan, Kun Gai, and Yang You. DiffusionART: Efficient, stable and scalable reinforcement of diffusion models. *arXiv preprint arXiv:2610.12572*, 2026.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. GRPO-Guard: Mitigating implicit over-optimization in flow matching via regulated clipping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Qingyan Wei, Guangzhao Li, Xiaobing Tu, Yinggui Wang, Xiantao Zhang, Jinkui Ren, Xiaohong Liu, and Linfeng Zhang. STEP-OPD: Rethinking output targets and internal dynamics in on-policy distillation for diffusion models. *arXiv preprint arXiv:2608.04887*, 2026.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, Linfeng Li, Shengqiong Wu, Leigang Qu, Feng Li, Wei Liu, Julian J. McAuley, and Tat-Seng Chua. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026a.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, and Tat-Seng Chua. DanceOPD: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026b.

A Implementation and Evaluation Details

A.1 Model and Optimization Settings

We initialize a fresh student LoRA adapter on SD3.5-Medium using Gaussian weights. The adapter attaches to the attention projections `to_q`, `to_k`, `to_v`, `to_out.0`, `add_q_proj`, `add_k_proj`, `add_v_proj`, and `to_add_out`. Only the student LoRA adapter is trained; the diffusion backbone, text encoders, VAE, and three released task specialist teachers remain frozen throughout training. Table A.1 summarizes our default training configuration.

Table A.1 Default DexOPD training configuration. One update corresponds to a single optimizer step.

Setting	Value
Image resolution	512×512
LoRA rank / scaling α	32 / 64
Optimizer / learning rate	AdamW / 3×10^{-4}
Adam coefficients (β_1, β_2)	$(0.9, 0.999)$
Adam ϵ / weight decay	10^{-8} / 10^{-4}
Gradient norm clipping	1.0
Sampling steps / supervised states	10 / 9
Samples per round / per update	144 / 72
Teacher extrapolation λ / auxiliary weight β	1.2 / 1.8
Candidate gains $\mathcal{G}$	$\{0, 0.05, 0.25, 0.5\}$
History EMA decay τ	0.995
Covariance RMS floor ε_C	10^{-8}
Training budget (updates / rounds)	1,000 / 500
Mixed precision	BF16 autocast
Student / history parameter precision	FP32
Checkpoint evaluation interval	100 updates

Classifier-Free Guidance. We apply classifier-free guidance as $v_{\text{cfg}} = v_{\text{uncond}} + w(v_{\text{cond}} - v_{\text{uncond}})$. During training, both the student and its historical reference use $w = 4.5$. Each task teacher and its corresponding base reference share the same task-specific guidance scale: $w = 1.0$ for GenEval and $w = 4.5$ for OCR and Aesthetic, evaluated at identical latents, timesteps, and text embeddings.

A.2 Collection and Update Schedule

At the start of round r, we freeze the student θ_r and historical reference H_r and collect 48 native rollouts per task (144 total). Each rollout starts from a Gaussian latent and executes ten deterministic Euler transitions from noise to data:

$$\mathcal{S}_j(x, v) = x + \Delta\sigma_j v, \qquad \Delta\sigma_j = \sigma_{j+1} - \sigma_j < 0. \tag{A.1}$$

We cache the native states, native velocities v_{old}, timesteps, noise scales σ_j, task identifiers, and prompt embeddings. States $j = 0, \ldots, 8$ receive supervision, while the tenth transition completes the clean endpoint latent for reward scoring.

For every prompt and initial noise latent, the native rollout endpoint directly supplies the $g = 0$ candidate. Each positive gain $g \in \{0.05, 0.25, 0.5\}$ integrates a full ten-step trajectory under the self-extrapolated velocity $v^{(g)} = (1+g)v_{\theta_r} - g v_{H_r}$, with constant g along the trajectory and both models queried at each intermediate state of that trajectory. The corresponding task reward model scores all four decoded endpoints (with the Aesthetic scorer used for the visual-preference branch).

We then compute and detach both the extrapolated teacher targets and auxiliary targets at the native states, caching them alongside the binary activity masks. Complete sample records are shuffled into two disjoint mini-batches of 72 samples, preserving exact task balance over each 144-sample round. Each optimizer update traverses its batch's nine supervised states once; the second update recomputes student predictions with the updated weights while reusing the round's cached targets. After both updates complete, the historical reference is refreshed via

$$H_{r+1} = 0.995\, H_r + 0.005\, \theta_{r+1}, \qquad H_0 = \theta_0, \tag{A.2}$$

where θ_{r+1} denotes the student adapter after the second update. Over 500 rounds (1,000 updates), each task contributes 24,000 native trajectories and four-candidate groups.

A.3 Endpoint Covariance, Targets, and Loss Reduction

For a candidate group of size $K = 4$ with endpoint latents $\{z^{(k)}\}_{k=0}^{K-1}$ and scalar rewards $\{r_k\}_{k=0}^{K-1}$, we compute the covariance tensor C and its RMS magnitude across all d latent elements:

$$C = \frac{1}{K}\sum_{k=0}^{K-1}(z^{(k)} - \bar{z})(r_k - \bar{r}), \qquad s_C = \sqrt{\frac{1}{d}\sum_{i=1}^{d} C_i^2}. \tag{A.3}$$

The binary activity mask a and normalized direction G are defined as

$$a = \mathbf{1}\left[\max_k r_k \neq \min_k r_k\right]\mathbf{1}\left[s_C > 10^{-8}\right], \qquad G = \begin{cases} C/s_C, & a = 1, \\ 0, & a = 0. \end{cases} \tag{A.4}$$

A single direction G and mask a are shared across all nine supervised states of the corresponding sample.

At each cached native state $x_{b,j}$, writing $v_{\text{proxy},q} = v_{T_q} + \lambda(v_{T_q} - v_B)$, the detached regression targets are

$$\begin{aligned} \mu_{T,b,j} &= \text{sg}[\mathcal{S}_j(x_{b,j}, v_{\text{proxy},q})], \\ \mu_{\text{aux},b,j} &= \text{sg}[\mathcal{S}_j(x_{b,j}, v_{\text{old},b,j} - G_b)]. \end{aligned} \tag{A.5}$$

Gradients flow strictly through the student predictions $\mu_{\theta,b,j} = \mathcal{S}_j(x_{b,j}, v_\theta(x_{b,j}, \sigma_j, c_b))$. For global mini-batch size $M = 72$, the empirical losses are

$$\begin{aligned} \mathcal{L}_{\text{OPD}} &= \frac{1}{9M}\sum_{b=1}^{M}\sum_{j=0}^{8}\tfrac{1}{2}\ \text{MSE}(\mu_{\theta,b,j}, \mu_{T,b,j}), \\ \mathcal{L}_{\text{ART}} &= \frac{1}{9M}\sum_{b=1}^{M}\sum_{j=0}^{8}\sigma_j a_b\ \text{MSE}(\mu_{\theta,b,j}, \mu_{\text{aux},b,j}), \\ \mathcal{L} &= \mathcal{L}_{\text{OPD}} + \beta\,\mathcal{L}_{\text{ART}}, \end{aligned} \tag{A.6}$$

where MSE averages over latent coordinates, and inactive groups ($a_b = 0$) contribute zero auxiliary loss while remaining in the batch denominator M.

A.4 Inference, Evaluation, and Prompt Splits

Whereas the lagging history H_r in Equation (A.2) generates training candidates, an evaluation EMA of the student adapter is used for single-trajectory test-time inference with no auxiliary branches. Following standard multi-task evaluation protocols, we report seven metrics spanning compositional alignment (GenEval), text rendering (OCR), and visual preference (Aesthetic, PickScore, ImageReward, CLIPScore, and HPSv2.1). We adopt the GenEval, OCR, and Pick-a-Pic training splits from DiffusionOPD (Li et al., 2026a). Table A.2 reports row counts and unique prompt counts for each split; GenEval generates four images per evaluation prompt and DrawBench generates five.

Table A.2 Training and evaluation prompt splits.

Split	File rows	Unique prompts
GenEval training	50,000	33,199
OCR training	19,653	19,649
Pick-a-Pic training	25,432	25,432
GenEval evaluation	2,212	553
OCR evaluation	1,018	1,018
DrawBench evaluation	1,000	200

A.5 Baseline Adaptations

All trained baselines share the SD3.5-Medium backbone and a 1,000-update training budget:

- **DiffusionOPD** collects 72 native trajectories and performs one optimizer update per round.
- **DreOPD** performs three updates of 24 samples per round, using an 8-bit per-channel velocity reference with extrapolation coefficient 1.25.
- **STEP-OPD*** collects 72 trajectories per update, supervises ten rollout states, and applies a 50-update warm-up with RCA $\varepsilon = 10^{-6}$.
- **Multi-task GRPO-Guard and DiffusionNFT** use one 24-sample group per task in each update. GRPO-Guard uses stochastic flow rollouts with learning rate 10^{-4}; DiffusionNFT uses deterministic DPM2 with learning rate 3×10^{-4}. The visual-preference reward equally weights PickScore, CLIPScore, and HPSv2.1.
- **Multi-task DiffusionART** collects six 24-sample groups per round with learning rate 3×10^{-4}, performing two updates per round for 500 rounds with all ten states supervised.

B Reward Aggregation and the Auxiliary Target

B.1 Pairwise Reward Contrasts

Let $z^{(k)}$ and r_k denote the endpoint latent and scalar reward of candidate k in a K-candidate group. The endpoint–reward covariance admits the exact pairwise identity

$$C = \frac{1}{K}\sum_{k=0}^{K-1}(z^{(k)} - \bar{z})(r_k - \bar{r}) = \frac{1}{K^2}\sum_{0\le k<\ell<K}(z^{(k)} - z^{(\ell)})(r_k - r_\ell), \tag{B.1}$$

since expanding the right-hand side yields $\frac{1}{K^2}[K\sum_k z^{(k)}r_k - (\sum_k z^{(k)})(\sum_k r_k)]$. Every pair (k,ℓ) contributes a vector pointing toward the higher-reward endpoint, weighted proportionally by their reward gap, including pairs among distinct positive gains.

B.2 Reward Scale Invariance and Activity

Let $s_C = \mathrm{RMS}(C)$. A candidate group is active ($a = 1$) whenever rewards are non-constant and $s_C > 10^{-8}$, ensuring that $G = C/s_C$ has unit RMS magnitude; inactive groups set $G = 0$. Under any positive affine reward rescaling $r'_k = \alpha r_k + b$ ($\alpha > 0$), both C and s_C scale identically by α, leaving G invariant so long as αs_C clears the numerical floor. Conversely, nonlinear transformations can alter relative gaps and rotate G. Uniform rewards, coincident endpoints, or exact pairwise cancellation each yield $C = 0$. At initialization ($H_0 = \theta_0$), all candidate trajectories coincide, so $G = 0$ cleanly defers auxiliary updates until extrapolated teacher supervision moves the student away from its initial history.

B.3 Displacement Sign, Scale, and Blended Regression Target

Under the noise-to-data Euler step $\mathcal{S}_j(x, v) = x + \Delta\sigma_j v$ with $\Delta\sigma_j = \sigma_{j+1} - \sigma_j < 0$, the auxiliary target relative to the cached transition mean $\mu_{\text{old},j} = \mathcal{S}_j(x_j, v_{\text{old},j})$ satisfies

$$\mu_{\text{aux},j} - \mu_{\text{old},j} = -\Delta\sigma_j G = |\Delta\sigma_j|\, G. \tag{B.2}$$

Hence, for any active group, the auxiliary target shifts strictly along $+G$ with RMS displacement $|\Delta\sigma_j|$. By linearity of $\mathcal{S}_j$, the transition-space loss corresponds to a weighted velocity-space regression:

$$\text{MSE}(\mu_{\theta,j}, \mu_{\text{aux},j}) = (\Delta\sigma_j)^2\ \text{MSE}(v_{\theta,j}, v_{\text{old},j} - G), \tag{B.3}$$

giving an effective auxiliary velocity weight of $\beta a\, \sigma_j (\Delta\sigma_j)^2$ across the noise schedule.

Writing $u = \mu_{\theta,j}$ and $w = \beta\sigma_j a$, completing the square at a single state reveals the exact blended regression target:

$$\begin{aligned} \tfrac{1}{2}\ \text{MSE}(u, \mu_{T,j}) + w\ \text{MSE}(u, \mu_{\text{aux},j}) &= \left(\tfrac{1}{2} + w\right) \text{MSE}(u, \mu_{*,j}) + c_j, \\ \mu_{*,j} &= \frac{\mu_{T,j} + 2w\, \mu_{\text{aux},j}}{1 + 2w}, \end{aligned} \tag{B.4}$$

where c_j is independent of u. When $a = 0$, the effective target reduces identically to the extrapolated teacher target $\mu_{T,j}$; when $a = 1$, larger β shifts $\mu_{*,j}$ smoothly toward $\mu_{\text{aux},j}$ while increasing its local weight. Thus, the candidate gains $\mathcal{G}$ shape the trajectories that induce the unit-RMS direction G, whereas β governs the relative pull of the resulting target. Across our default run (Appendix C), GenEval relies solely on the teacher target in 90.50% of groups, whereas Aesthetic uses the blended target in 99.80% of groups.

C Online Candidate Diagnostics

Across the default 500-round training run, we log all 24,000 four-candidate groups per task. Table C.1 compares each positive self-extrapolation gain against its same-noise native endpoint ($g = 0$) using online training-prompt rewards.

Table C.1 Pairwise outcomes of positive-gain candidates against the same-noise native endpoint. Each row aggregates 24,000 online comparisons across the default run; $\Delta r = r_g - r_0$ denotes the mean signed reward difference.

Task	g	Δr	Win (%)	Tie (%)	Loss (%)
GenEval	0.05	−0.0003	1.4708	96.9750	1.5542
	0.25	−0.0010	2.7375	94.4667	2.7958
	0.50	−0.0065	3.1583	92.6292	4.2125
OCR	0.05	0.0002	4.6833	90.5875	4.7292
	0.25	0.0012	7.9250	83.7125	8.3625
	0.50	0.0004	9.4000	79.8667	10.7333
Aesthetic	0.05	−0.0021	49.5333	0.2000	50.2667
	0.25	−0.0194	45.5083	0.2000	54.2917
	0.50	−0.0652	39.8667	0.2000	59.9333

On GenEval, positive-gain candidates tie the native endpoint in 92.63–96.98% of comparisons because GenEval's object-detection verifier produces discrete scores that often coincide across nearby same-noise samples. On OCR, mean signed reward differences are positive across all three gains, while pairwise losses outnumber wins due to asymmetric score jumps. On Aesthetic, increasing g from 0.05 to 0.50 shifts the pairwise win rate from 49.53% to 39.87%, with negative mean reward differences across all three gains for both GenEval and Aesthetic. Because covariance aggregation (Equation (B.1)) weights both higher- and lower-reward pairs symmetrically, this reward spread translates directly into a consistent corrective direction. Overall full-run activity rates are 9.50% for GenEval, 26.44% for OCR, and 99.80% for Aesthetic.

Activity Dynamics Over Training. Comparing the first 200 updates against the final 200 updates (Figure C.1), mean native reward climbs from 0.8632 to 0.9636 on GenEval and from 0.9340 to 0.9721 on OCR. As native predictions saturate on easier prompts, candidate activity correspondingly tapers from 20.52% to 4.94% on GenEval and from 32.15% to 24.21% on OCR, while Aesthetic activity stays near 100% throughout training because continuous visual-preference scores almost always separate distinct candidates.

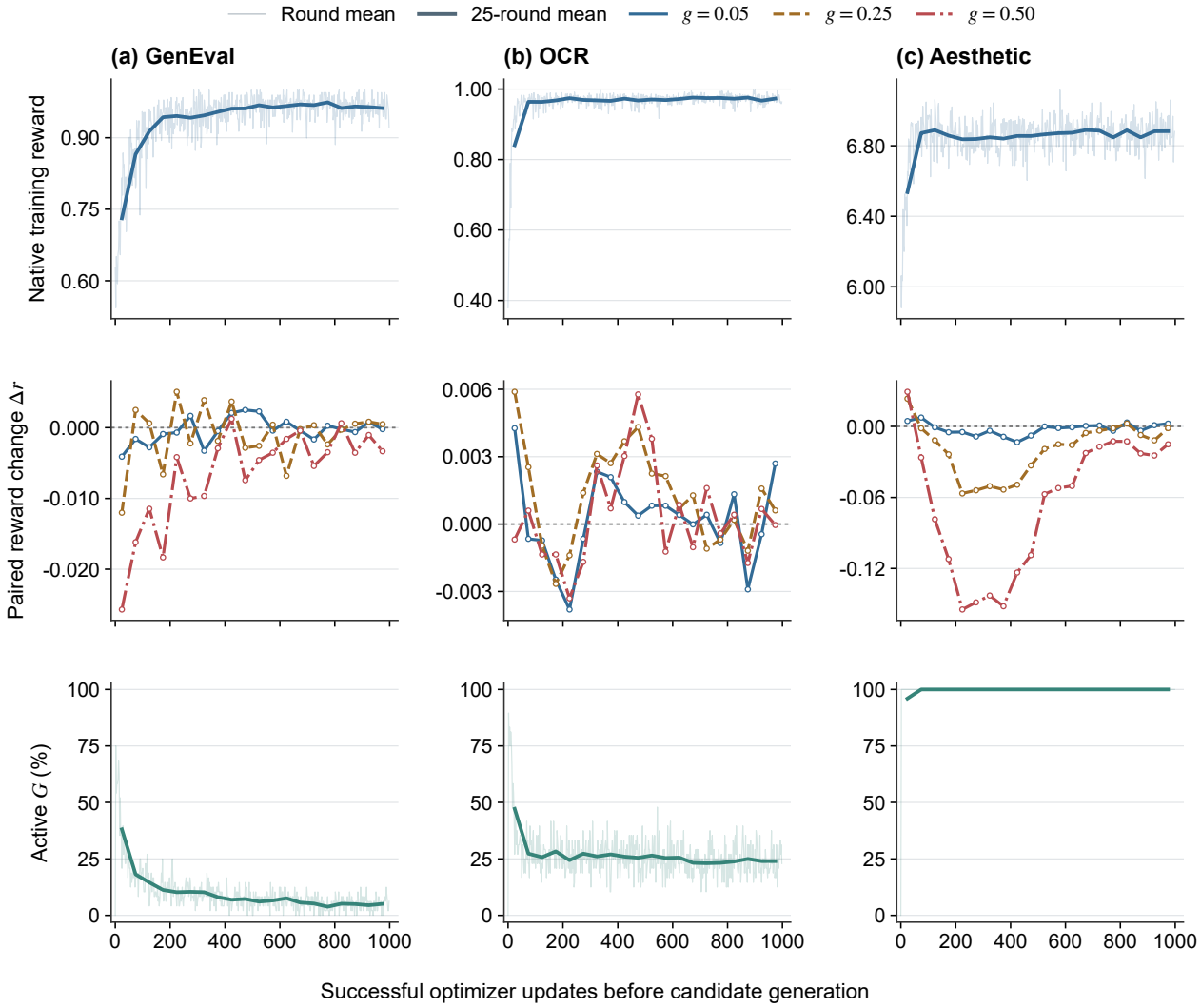


Figure C.1 Online candidate rewards and auxiliary activity across training. Rows display native rollout reward, mean candidate reward difference $\Delta r = r_g - r_0$, and the fraction of groups with active direction G. Thin traces indicate per-round means; bold curves average 25-round windows (1,200 groups per task). The horizontal axis indexes optimizer updates prior to each collection round.

D Complete Auxiliary-Weight Ablation

Table D.1 extends Table 3 to the full 11-point sweep over $\beta \in [0.0, 2.0]$ with $\lambda = 1.2$ held fixed. Our default setting ($\beta = 1.8$) achieves the highest DrawBench Aesthetic score (6.4875) while remaining within 0.0007 of the GenEval peak ($\beta = 0.4$) and within 0.0005 of the OCR peak ($\beta = 2.0$). By contrast, the OCR-leading setting ($\beta = 2.0$) trails our default by 0.0149 on GenEval and 0.0147 on Aesthetic.

Every positive weight improves Aesthetic and ImageReward relative to $\beta = 0$: at $\beta = 1.8$, the gains over $\beta = 0$ are +0.0067 on GenEval, +0.0107 on OCR, and +0.0465 on Aesthetic. The optimal auxiliary weight is

Table D.1 Complete auxiliary-weight sweep at 1,000 updates. All runs fix $\lambda = 1.2$. Bold and underline mark the best and second-best values for each metric; the highlighted row denotes the default $\beta = 1.8$. Higher is better for all metrics.

β	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
0.0	0.9669	0.9408	6.4410	0.9185	1.4957	<u>0.2944</u>	<u>0.3527</u>
0.2	0.9715	0.9356	6.4511	0.9185	1.5086	0.2922	0.3485
0.4	**0.9743**	0.9414	6.4533	0.9190	1.5036	0.2911	0.3478
0.6	0.9681	0.9460	6.4629	0.9187	1.5023	0.2912	0.3484
0.8	0.9647	0.9452	6.4689	0.9192	1.5045	0.2919	0.3477
1.0	0.9698	0.9478	6.4594	0.9195	1.5124	**0.2952**	**0.3532**
1.2	0.9717	0.9500	6.4718	<u>0.9198</u>	1.5123	0.2919	0.3481
1.4	0.9702	0.9458	6.4448	0.9192	**1.5177**	0.2921	0.3482
1.6	0.9526	0.9437	6.4622	**0.9201**	1.5118	0.2919	0.3470
1.8	<u>0.9736</u>	<u>0.9515</u>	**6.4875**	0.9190	<u>1.5127</u>	0.2914	0.3484
2.0	0.9587	**0.9520**	<u>6.4728</u>	0.9187	1.5108	0.2920	0.3471

metric-dependent across secondary preference scores: CLIPScore and HPSv2.1 peak at $\beta = 1.0$, ImageReward peaks at $\beta = 1.4$, and PickScore peaks at $\beta = 1.6$.

E Complete Teacher Inference Extrapolation Sweep

This appendix extends Figure 6 and Table 4 by evaluating standalone inference-time extrapolation $v_T + \lambda(v_T - v_B)$ across all 11 values $\lambda \in \{0.0, 0.2, \ldots, 2.0\}$ for each specialist teacher (Figure E.1).

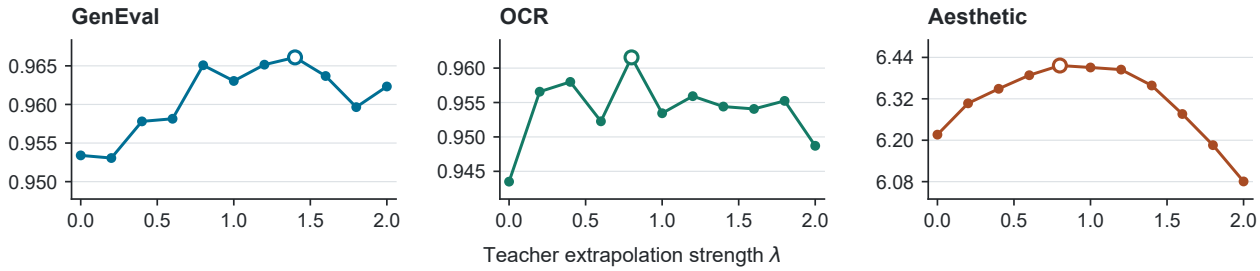


Figure E.1 On-task performance across the full native teacher extrapolation sweep. Open circles mark each specialist teacher's peak score on its target benchmark.

GenEval Teacher. As detailed in Table E.1, the GenEval teacher's target reward climbs from 0.9534 at $\lambda = 0.0$ to a peak of 0.9661 at $\lambda = 1.4$, remaining within 0.0065 of this maximum throughout $\lambda \in [0.8, 2.0]$.

OCR Teacher. As shown in Table E.2, the OCR specialist peaks at 0.9616 at $\lambda = 0.8$ (a +0.0181 gain over its native score of 0.9435 at $\lambda = 0.0$), before receding to 0.9487 at $\lambda = 2.0$.

Aesthetic Teacher. In Table E.3, the Aesthetic teacher improves by +0.2002 to reach 6.4164 at $\lambda = 0.8$, stays above 6.40 through $\lambda = 1.2$, and drops sharply to 6.0806 at $\lambda = 2.0$. Among auxiliary preference scores, ImageReward peaks at $\lambda = 0.6$ (1.5077), whereas PickScore declines steadily for $\lambda > 0$. Its off-task GenEval score rises briefly to 0.5288 at $\lambda = 0.2$ (0.5011 at $\lambda = 0.4$) before falling to 0.2288 at $\lambda = 2.0$.

Table E.1 Full GenEval teacher inference sweep. Bold marks the on-task maximum.

λ	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
0.0	0.9534	0.6515	5.2635	0.8513	0.9277	0.2869	0.2673
0.2	0.9531	0.4157	5.2942	0.8446	0.5511	0.2745	0.2463
0.4	0.9578	0.4302	5.2781	0.8439	0.5093	0.2739	0.2458
0.6	0.9581	0.4316	5.2622	0.8424	0.4636	0.2732	0.2443
0.8	0.9651	0.4559	5.2781	0.8420	0.4716	0.2717	0.2452
1.0	0.9630	0.4775	5.2512	0.8410	0.4545	0.2709	0.2439
1.2	0.9652	0.4714	5.2611	0.8401	0.4643	0.2710	0.2429
1.4	**0.9661**	0.4985	5.2525	0.8387	0.4002	0.2681	0.2418
1.6	0.9637	0.4733	5.2501	0.8375	0.4041	0.2672	0.2415
1.8	0.9597	0.4818	5.2543	0.8364	0.3683	0.2648	0.2400
2.0	0.9623	0.4872	5.2526	0.8358	0.3536	0.2645	0.2387

Table E.2 Full OCR teacher inference sweep. Bold marks the on-task maximum.

λ	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
0.0	0.6617	0.9435	5.2605	0.8580	0.8972	0.2909	0.2732
0.2	0.6214	0.9566	5.2438	0.8391	0.6339	0.2912	0.2508
0.4	0.5914	0.9580	5.1923	0.8357	0.5848	0.2891	0.2456
0.6	0.5765	0.9523	5.1718	0.8314	0.5108	0.2879	0.2423
0.8	0.5613	**0.9616**	5.1091	0.8262	0.4202	0.2855	0.2370
1.0	0.5574	0.9534	5.0735	0.8223	0.3461	0.2816	0.2333
1.2	0.5145	0.9559	5.0281	0.8166	0.2917	0.2807	0.2299
1.4	0.5004	0.9544	4.9383	0.8101	0.0857	0.2765	0.2220
1.6	0.4695	0.9541	4.8835	0.8055	−0.0227	0.2711	0.2177
1.8	0.4459	0.9552	4.8332	0.7994	−0.1318	0.2665	0.2133
2.0	0.4276	0.9487	4.7608	0.7959	−0.2367	0.2646	0.2095

Table E.3 Full Aesthetic teacher inference sweep. Bold marks the on-task maximum.

λ	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
0.0	0.4791	0.5881	6.2162	0.9233	1.5016	0.2960	0.3452
0.2	0.5288	0.6173	6.3069	0.9222	1.4916	0.2962	0.3425
0.4	0.5011	0.6017	6.3491	0.9228	1.5071	0.2959	0.3444
0.6	0.4711	0.5949	6.3883	0.9204	1.5077	0.2941	0.3446
0.8	0.4499	0.5686	**6.4164**	0.9170	1.4932	0.2940	0.3440
1.0	0.4418	0.5491	6.4108	0.9128	1.4785	0.2925	0.3427
1.2	0.3903	0.5269	6.4047	0.9068	1.4589	0.2928	0.3398
1.4	0.3926	0.5126	6.3586	0.9007	1.4189	0.2917	0.3364
1.6	0.3331	0.4981	6.2760	0.8935	1.3907	0.2924	0.3307
1.8	0.2851	0.5207	6.1857	0.8847	1.3369	0.2908	0.3240
2.0	0.2288	0.4858	6.0806	0.8759	1.2604	0.2891	0.3160

Cross-Task Responses. The three specialists exhibit distinct off-task trade-offs under extrapolation (Figure E.2). For the GenEval teacher, OCR falls from 0.6515 at $\lambda = 0.0$ to 0.4157 at $\lambda = 0.2$ before partially recovering to 0.4985 at the GenEval peak, while ImageReward drops steadily from 0.9277 to 0.3536 at $\lambda = 2.0$. For the OCR teacher, GenEval drops monotonically from 0.6617 at $\lambda = 0.0$ to 0.5613 at the OCR peak ($\lambda = 0.8$) and 0.4276 at $\lambda = 2.0$; Aesthetic declines from 5.2605 at $\lambda = 0.0$ to 5.1091 at $\lambda = 0.8$ and 4.7608 at $\lambda = 2.0$, and ImageReward turns negative at $\lambda = 1.6$. Similarly, the Aesthetic teacher's GenEval score falls from 0.4791 at $\lambda = 0.0$ to 0.2288 at $\lambda = 2.0$.

At our training default of $\lambda = 1.2$, the GenEval and Aesthetic teachers remain within 0.0009 and 0.0117 of their respective on-task maxima, while the OCR teacher scores 0.9559 (0.0057 below its peak). As shown in Table 2, retaining teacher extrapolation at $\lambda = 1.2$ during distillation improves the unified student by +0.0221 on GenEval, +0.0158 on OCR, and +0.1945 on Aesthetic over $\lambda = 0$.

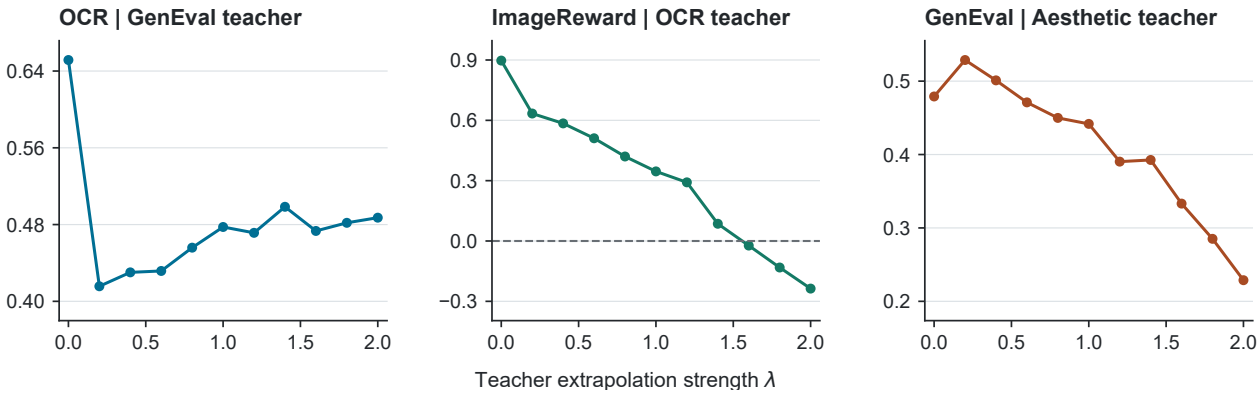


Figure E.2 Selected off-task degradation curves under teacher extrapolation. Each panel tracks an off-task metric during one specialist's λ sweep. The dashed horizontal line marks zero ImageReward.

F Small-Batch Efficiency and Training Time

F.1 Small-Batch Checkpoint Progression

To examine compute–performance scaling, we train a small-batch DexOPD variant (batch size 24) starting from a fresh student adapter, collecting 48 trajectories per round (16 per task) while keeping all optimizer, rollout, supervision, and extrapolation hyperparameters identical to Appendix A. Figure F.1 and Table F.1 report the evaluation EMA checkpoints every 100 updates.

Between updates 100 and 1,000, GenEval reward increases by +0.1138 and Aesthetic by +0.6085; OCR reaches 0.9485 by update 200 and peaks at 0.9488 at update 1,000. Secondary preference metrics peak near the end of training (PickScore at update 800; ImageReward and HPSv2.1 at update 900). At update 1,000, batch 24 trails the default batch-72 model by only 0.0056 on GenEval, 0.0027 on OCR, and 0.0018 on Aesthetic, despite using one-third as many native trajectories—and still surpasses all three OPD baselines in Table 1 on every primary task metric.

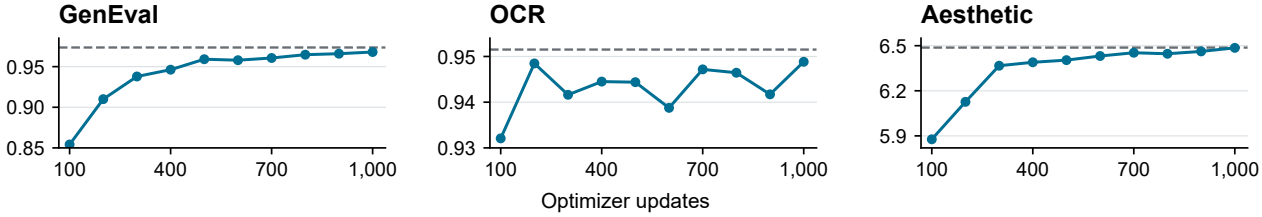


Figure F.1 Learning curves for small-batch DexOPD (batch size 24). Horizontal dashed lines denote the default batch-72 endpoints at update 1,000.

Table F.1 Seven-metric checkpoint progression for small-batch DexOPD (batch size 24). The bottom row lists the default batch-72 endpoint at 1,000 updates for reference. Bold marks the best batch-24 score on each primary target metric; higher is better for all metrics.

Batch	Updates	GenEval	OCR	Aesthetic	PickScore	ImageReward	CLIPScore	HPSv2.1
24	100	0.8542	0.9321	5.8772	0.8999	1.3730	0.2965	0.3147
24	200	0.9100	0.9485	6.1263	0.9069	1.4773	0.2957	0.3325
24	300	0.9379	0.9416	6.3668	0.9140	1.5182	0.2945	0.3423
24	400	0.9462	0.9445	6.3896	0.9146	1.5027	0.2910	0.3442
24	500	0.9591	0.9444	6.4040	0.9174	1.5090	0.2931	0.3443
24	600	0.9579	0.9388	6.4314	0.9183	1.4958	0.2935	0.3455
24	700	0.9606	0.9472	6.4532	0.9192	1.5072	0.2927	0.3473
24	800	0.9647	0.9464	6.4464	0.9200	1.5080	0.2930	0.3464
24	900	0.9659	0.9417	6.4620	0.9198	1.5188	0.2929	0.3480
24	1,000	**0.9680**	**0.9488**	**6.4857**	0.9192	1.5133	0.2917	0.3477
72	1,000	0.9736	0.9515	6.4875	0.9190	1.5127	0.2914	0.3484

F.2 Wall-Clock Training Time

Table F.2 compares average wall-clock training time per optimizer update on eight H20 GPUs alongside final target-task performance. Small-batch DexOPD averages 25.59 seconds per update—virtually identical to standard single-trajectory DiffusionOPD (25.76 seconds) and faster than STEP-OPD* (32.53 seconds)—while retaining nearly the full quality of batch-72 DexOPD and outperforming every OPD baseline on all three target metrics.

Table F.2 Training time and target-task performance. Average wall-clock time per optimizer update is measured on eight H20 GPUs. Batch indicates the number of native samples per update. Bold and underline denote the best and second-best target scores among the listed methods. *: paper-based reproduction.

Method	Batch	GenEval ↑	OCR ↑	Aesthetic ↑	s/update ↓
DiffusionOPD	72	0.9543	0.9371	6.1820	25.76
DreOPD	24	0.9622	0.9354	6.2182	9.73
STEP-OPD*	72	0.9562	0.9378	6.2813	32.53
DexOPD	72	**0.9736**	**0.9515**	**6.4875**	68.18
DexOPD (small batch)	24	<u>0.9680</u>	<u>0.9488</u>	<u>6.4857</u>	25.59

G Additional Generation Examples

We present 165 additional outputs generated by our unified DexOPD student. The 77 GenEval examples (Figures G.1 to G.6) demonstrate accurate object counting, unconventional color–object bindings, and fine-grained spatial relations. The 88 OCR examples (Figures G.7 to G.12) showcase legible text rendering across diverse surfaces including ceramic mugs, neon signs, storefronts, and fabric banners.

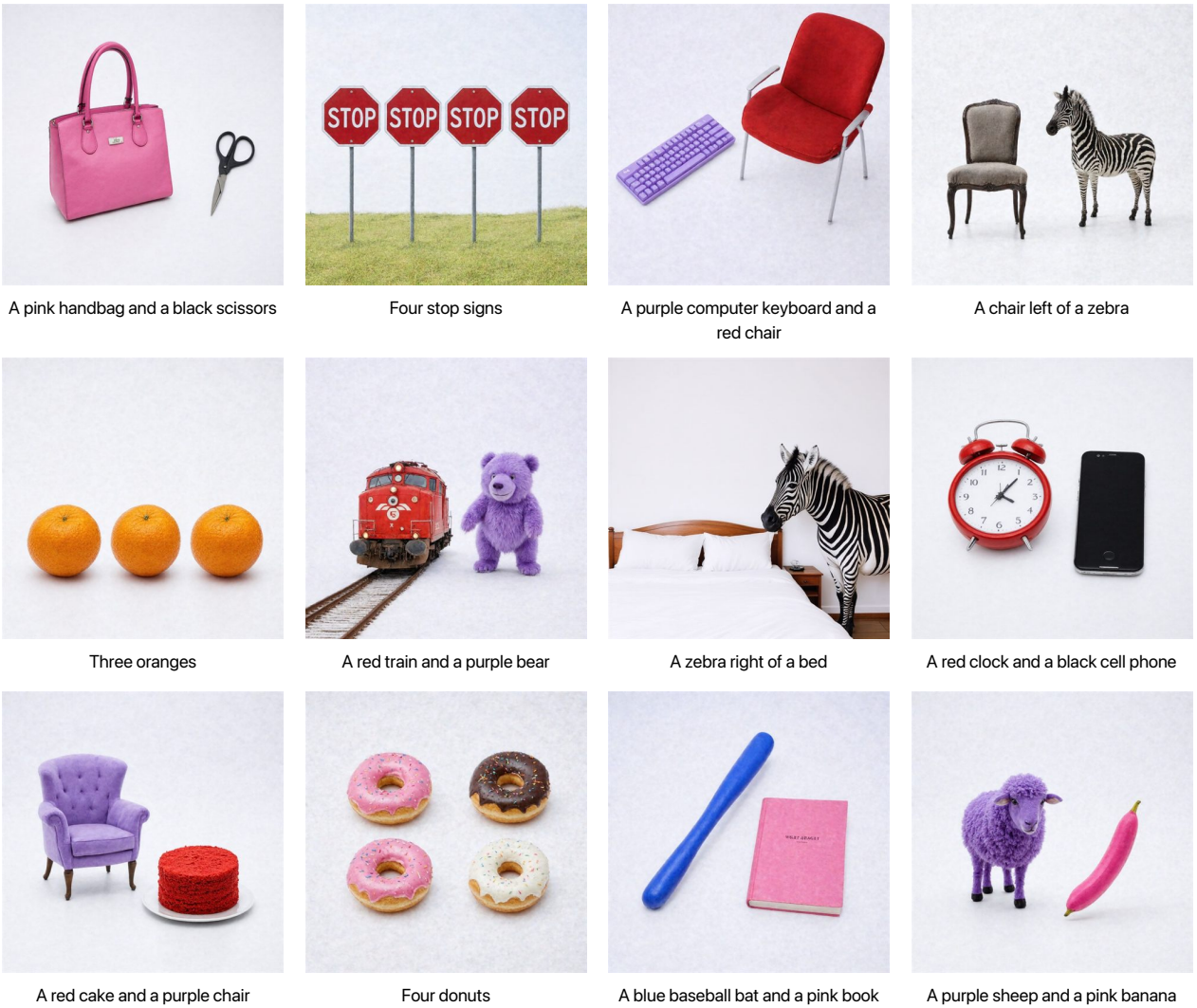


Figure G.1 Selected GenEval outputs from the unified DexOPD student (1/6). Labels summarize the requested objects, counts, colors, or spatial relations.

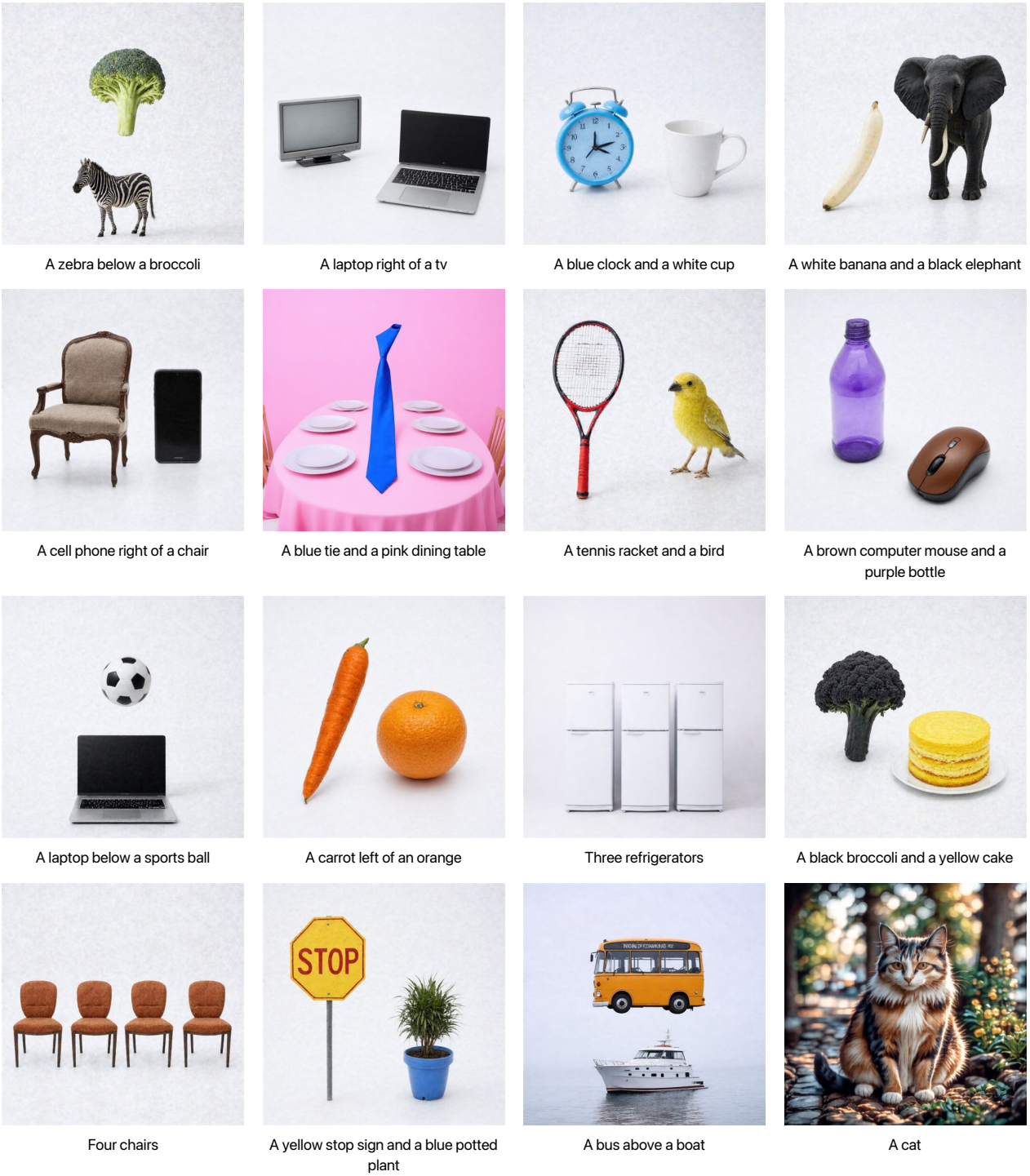


Figure G.2 Selected GenEval outputs from the unified DexOPD student (2/6). Labels summarize the requested objects, counts, colors, or spatial relations.

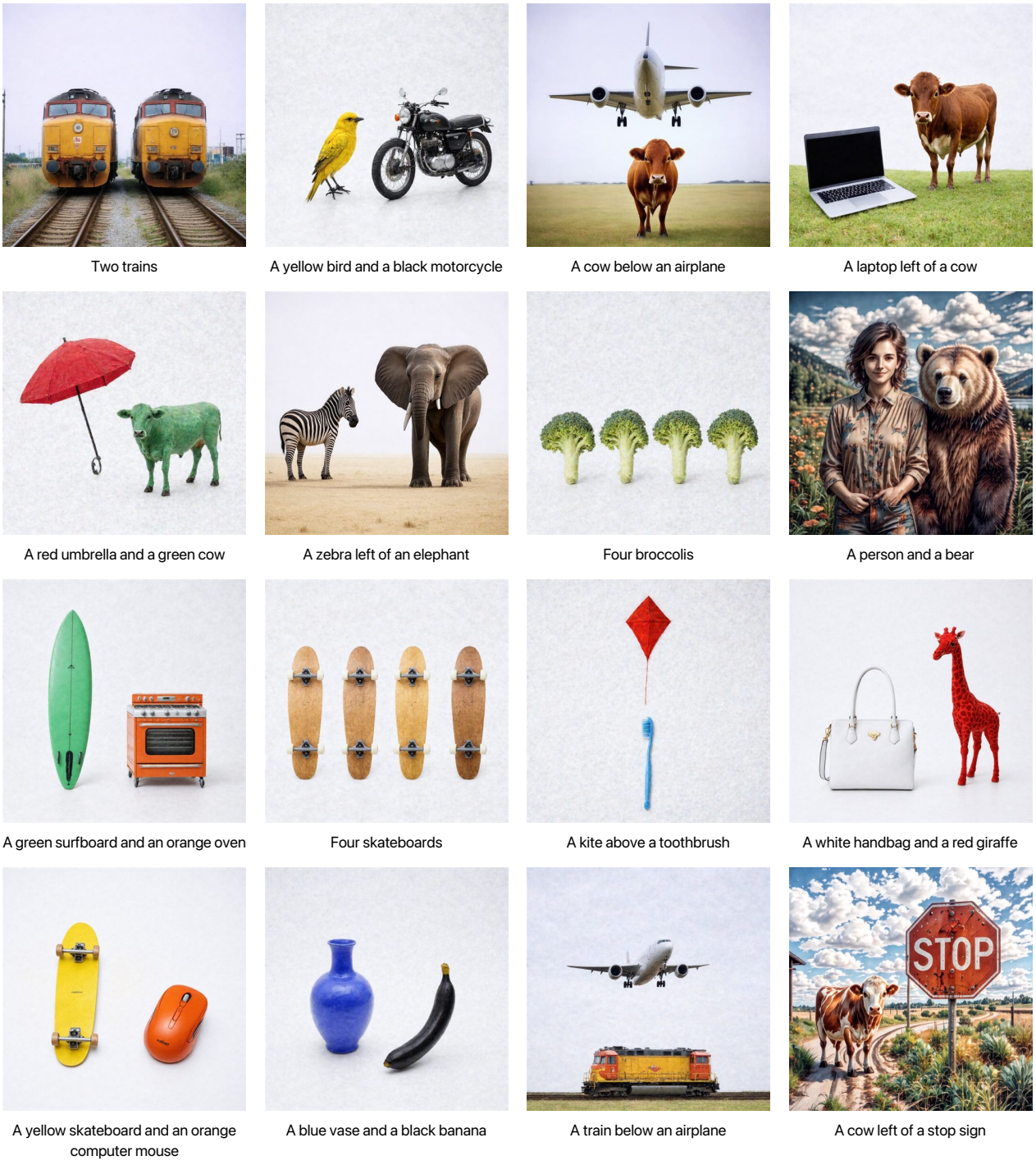


Figure G.3 Selected GenEval outputs from the unified DexOPD student (3/6). Labels summarize the requested objects, counts, colors, or spatial relations.



A blue cow and a black computer keyboard | A dining table and a bear | Three birds | A potted plant and a donut



A purple elephant and a brown sports ball | A yellow sports ball and a green boat | A blue fire hydrant | A suitcase right of a boat



A red apple | An elephant below a horse | A brown knife and a blue donut | A tie and a broccoli

Figure G.4 Selected GenEval outputs from the unified DexOPD student (4/6). Labels summarize the requested objects, counts, colors, or spatial relations.



Three buses | Three books | Three snowboards | A yellow computer keyboard and a black sink



A hair drier and a bear | Three kites | A dog above a cow | A green bus



An orange snowboard and a green cat | Two teddy bears | Three fire hydrants | A hot dog right of a skateboard

Figure G.5 Selected GenEval outputs from the unified DexOPD student (5/6). Labels summarize the requested objects, counts, colors, or spatial relations.

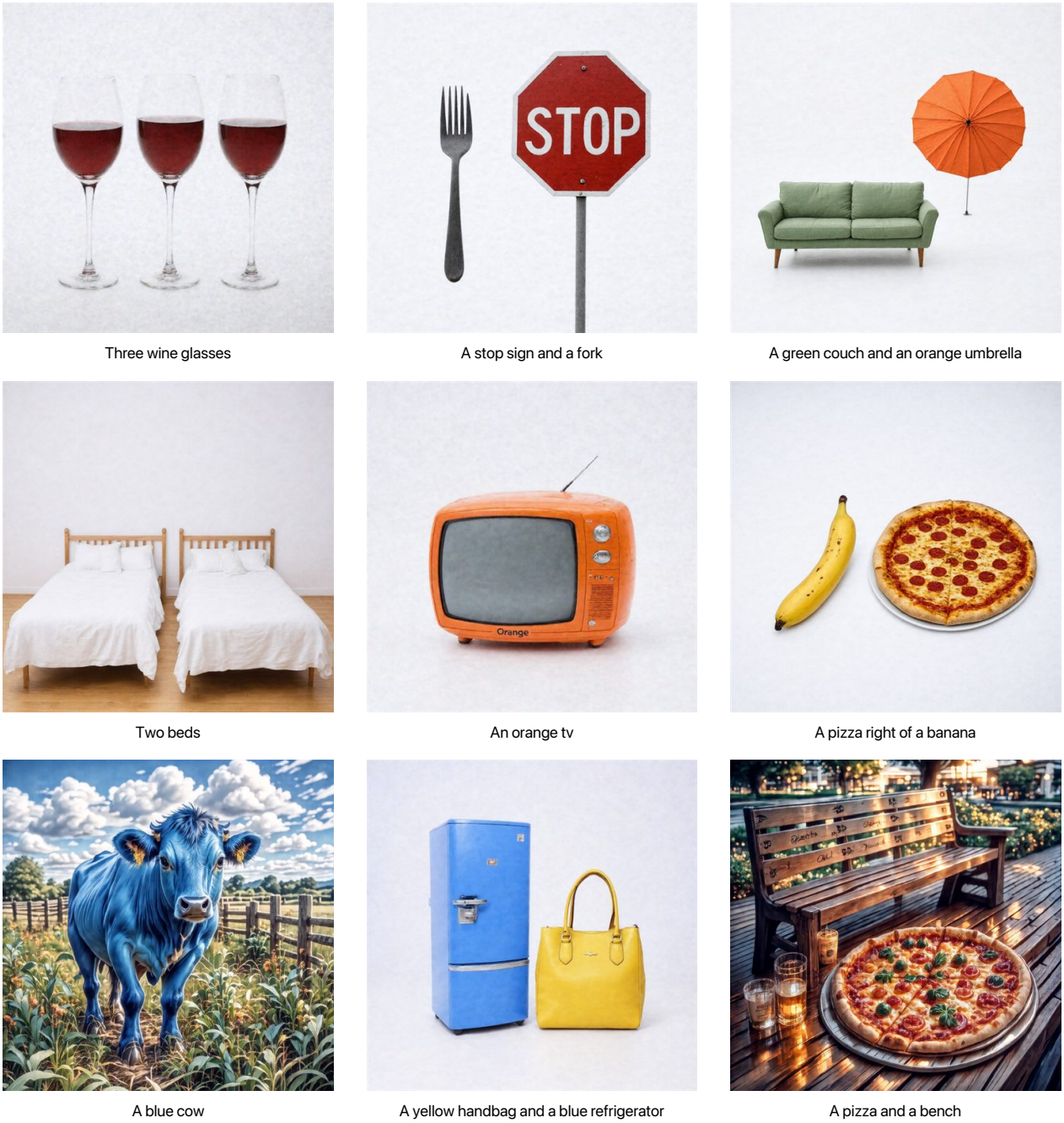


Figure G.6 Selected GenEval outputs from the unified DexOPD student (6/6). Labels summarize the requested objects, counts, colors, or spatial relations.

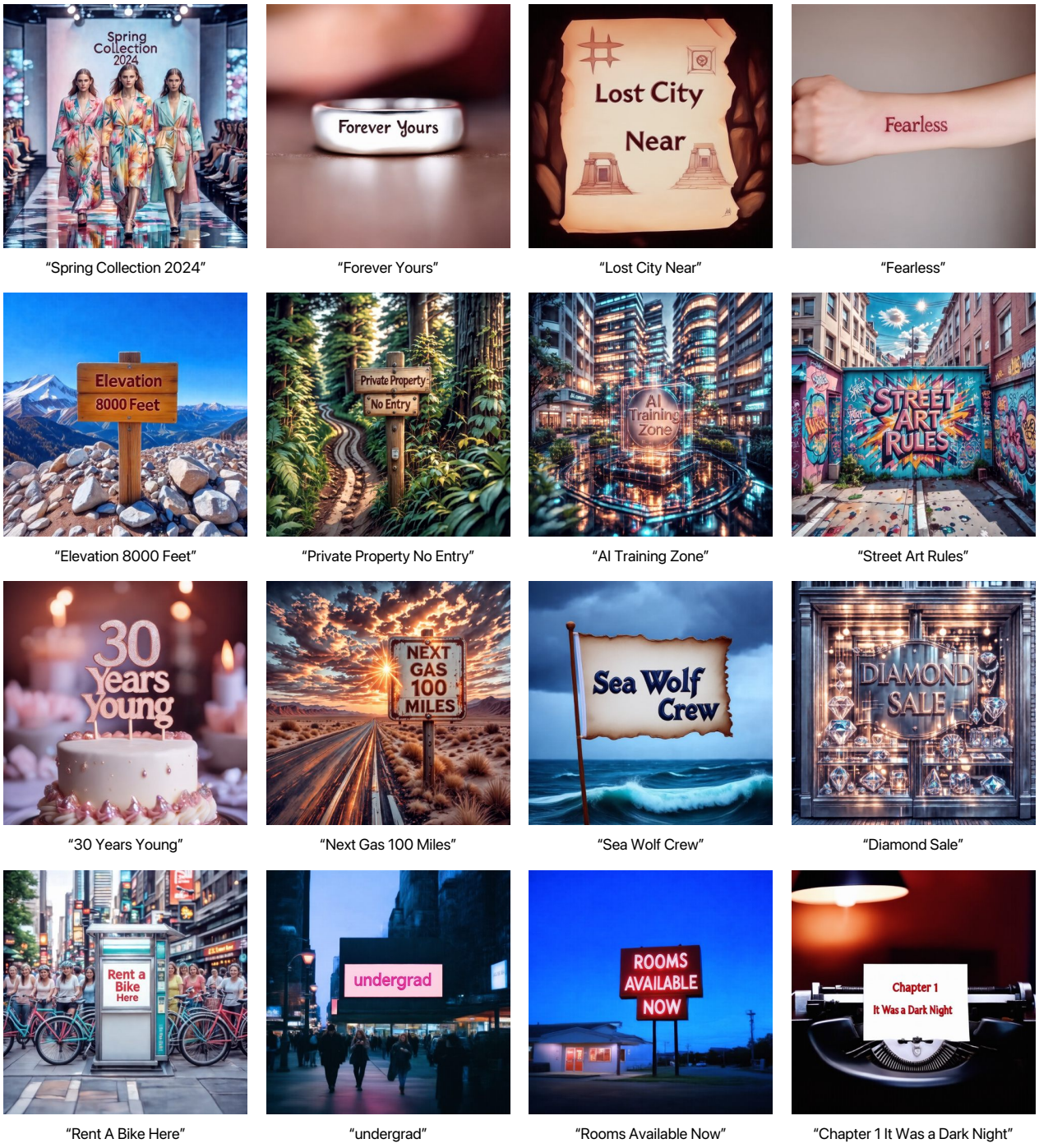


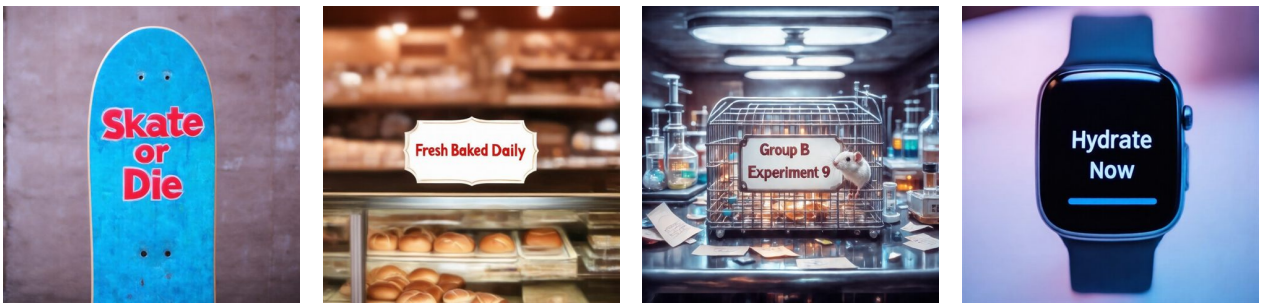
Figure G.7 Selected OCR outputs from the unified DexOPD student (1/6). Labels indicate the requested target text.



"Campfire Brew" "Sample XZ42" "Life's a Beach 2024" "Burn After Reading"



"Handle With Care" "Order in the Court" "Water Weekly with Stardust" "Now Serving 42"



"Skate or Die" "Fresh Baked Daily" "Group B Experiment 9" "Hydrate Now"



"Click Here for Discount" "Electric Vehicles Here" "Break Glass for Drama" "The Cake Was a Lie"

Figure G.8 Selected OCR outputs from the unified DexOPD student (2/6). Labels indicate the requested target text.

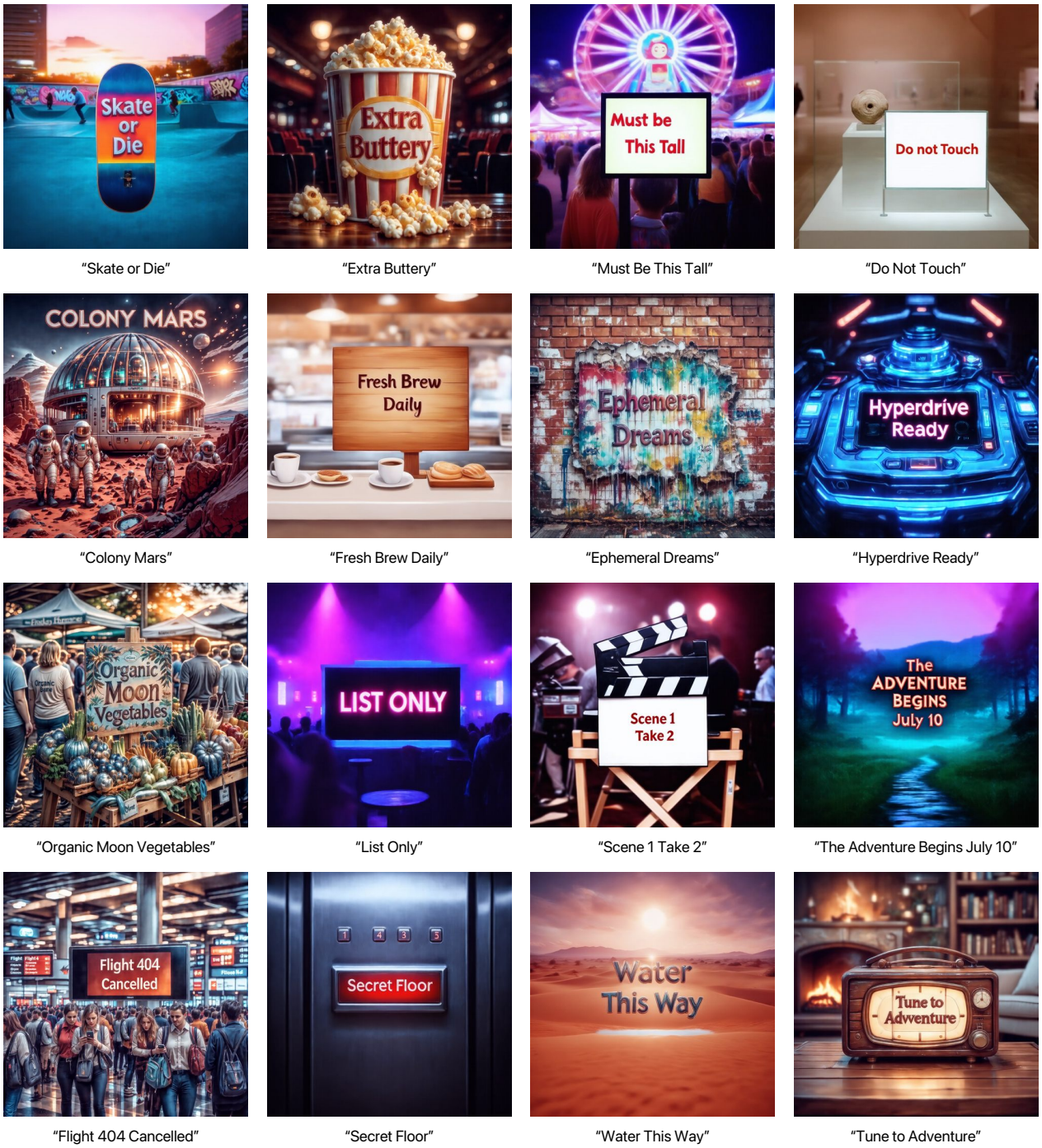


Figure G.9 Selected OCR outputs from the unified DexOPD student (3/6). Labels indicate the requested target text.

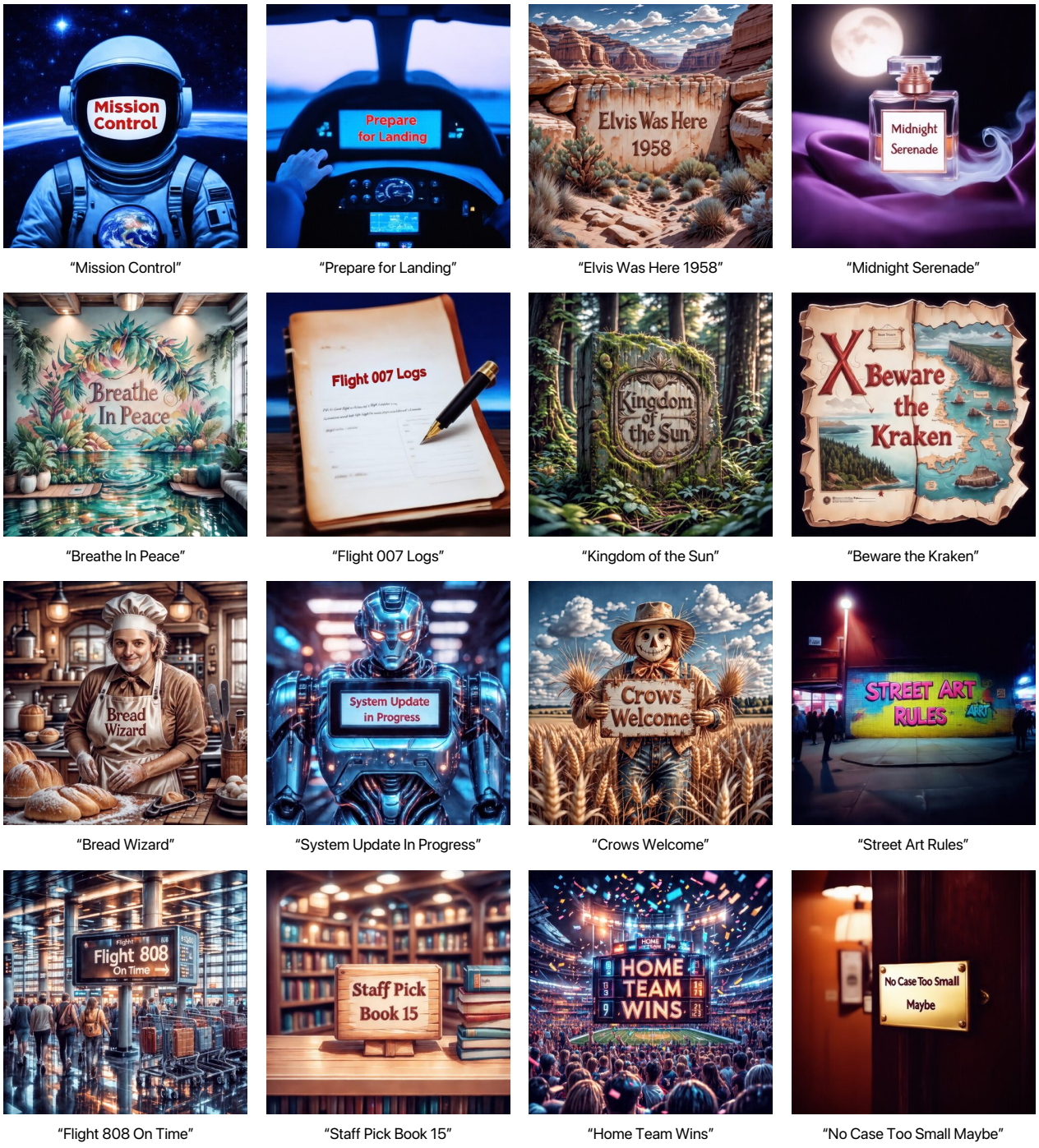


Figure G.10 Selected OCR outputs from the unified DexOPD student (4/6). Labels indicate the requested target text.

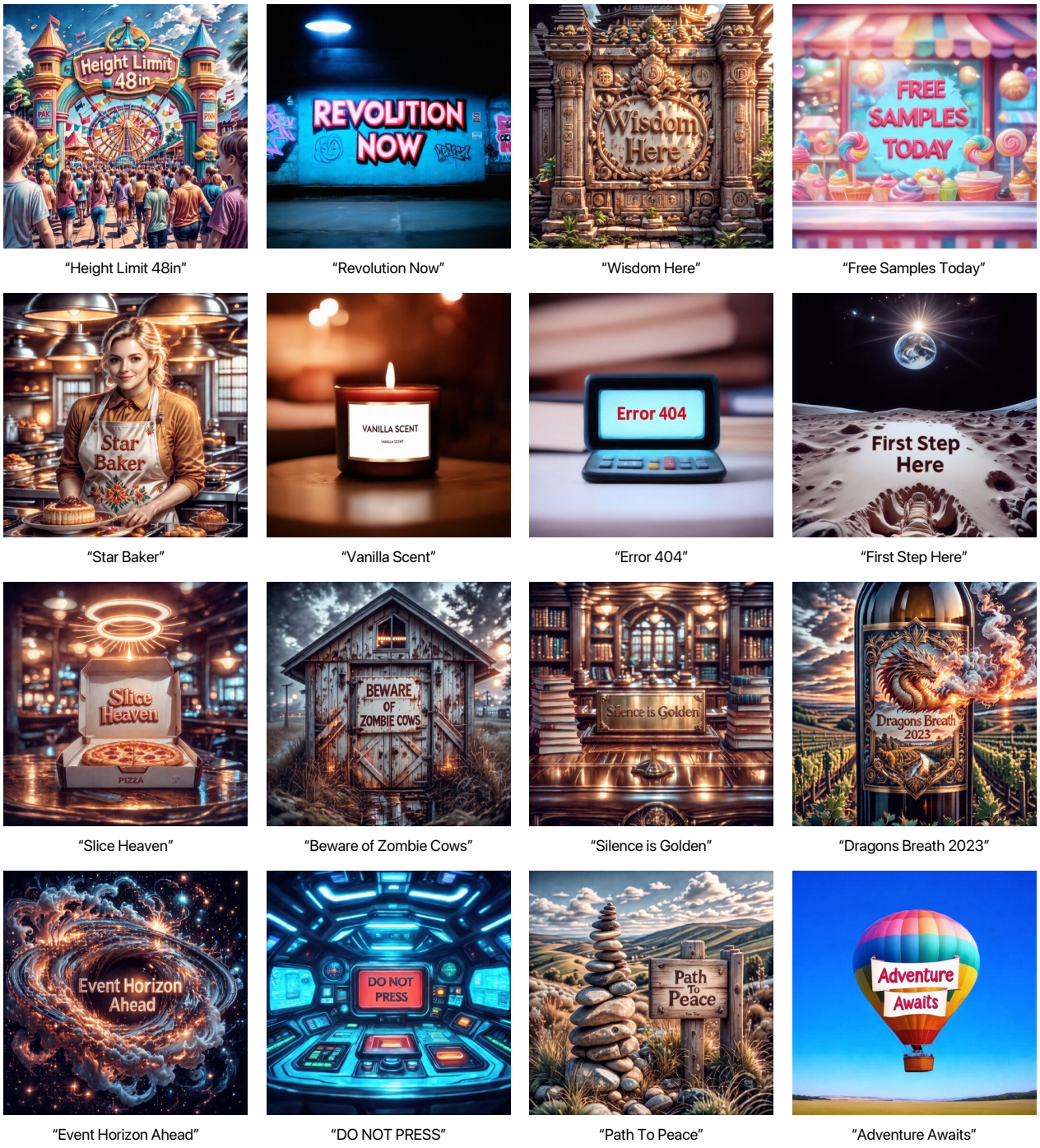


Figure G.11 Selected OCR outputs from the unified DexOPD student (5/6). Labels indicate the requested target text.



"Mad Genius" "Now Showing Action 5" "Festival Access 2024" "Live Free"



"World's Best Deli" "Established 1890" "Kingdom of Zorath" "0 APR Financing"

Figure G.12 Selected OCR outputs from the unified DexOPD student (6/6). Labels indicate the requested target text.