DreOPD: Degraded-Reference Extrapolative On-Policy Distillation for Flow-matching Models

Anonymous authors
Paper under double-blind review



Figure 1: **Left:** Representative images generated by our method across diverse subjects, scenes, and text-rendering scenarios. **Right:** Our method achieves the highest average performance across the evaluated reward metrics.

Abstract

Flow-matching models are now a mainstream method to image generation, but its adaptation to diverse downstream scenarios typically relies on post-training, which may cause conflicts among task-specific optimization objectives. Reinforcement learning enables direct optimization of task-specific rewards beyond the original models, yet trajectory-level optimization may incur high-variance gradients and cross-task interference. On-policy distillation (OPD) offers dense and stable supervision on student rollouts, but conventional teacher matching remains imitation-based. We propose **DreOPD**, a **D**egraded-**r**eference **e**xtrapolative **OPD** method for flow-matching models that bridges these two paradigms. Our **DreOPD** converts implicit reward extrapolation into closed-form velocity regression, enabling extrapolative post-training with the stability of OPD. It further uses a mildly degraded reference to strengthen the teacher-reference contrast, yielding a clearer extrapolation direction. Experiments on single- and multi-teacher settings show that **DreOPD** outperforms OPD and multi-task RL baselines in average performance, while surpassing specialized teachers on most metrics.

1 Introduction

Flow-matching and diffusion models have emerged as leading paradigms for text-to-image generation (Lipman et al., 2022; Cao et al., 2025). Despite their strong general capabilities, practical deployment often requires post-training for diverse objectives, such as prompt following, text rendering, aesthetics, and human preference (Li et al., 2026a; Lyu et al., 2026). These objectives may exhibit implicit conflicts, making it challenging for a single model to improve multiple capabilities simultaneously (Yu et al., 2020; Hu et al., 2024). Meanwhile, obtaining a strong task-specific teacher requires costly alignment, whereas weaker model variants are readily available through early training checkpoints or controlled degradation. Yet these weaker models are typically discarded rather than exploited as useful signals. This motivates a practical question: *how can a student consolidate multiple strong teachers, leverage readily available weaker models, and improve beyond the teachers themselves?*

Existing approaches expose a fundamental trade-off. Reinforcement learning (RL) (Black et al., 2024; Wallace et al., 2024; Li et al., 2025; Xue et al., 2025a; Wang et al., 2026; Liu et al., 2026)

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.

directly optimize task rewards and can potentially improve beyond existing models. However, trajectory-level supervision may suffer from high variance, reward exploitation, and interference among heterogeneous objectives. On-policy distillation (OPD) instead provides dense and stable supervision by matching teachers on student-generated states (Li et al., 2026b; Fang et al., 2026; Zhou et al., 2026). However, standard OPD is imitation-based: each teacher remains the pointwise target, causing a shared multi-task student to interpolate among specialized teachers rather than systematically exceed them. Thus, RL permits extrapolation but is difficult to optimize, whereas OPD is stable but teacher-bounded.

Reward extrapolation for autoregressive language models (Yang et al., 2026) offers a route between these two paradigms. It interprets the teacher-reference log-likelihood ratio as an implicit reward and amplifies this contrast beyond teacher matching. Directly transferring this principle to flow-matching models is nontrivial because such models predict continuous velocity fields rather than categorical token distributions (Nie et al., 2026). A trajectory-level implementation would require likelihood-ratio estimation and credit assignment across continuous denoising steps (Li et al., 2026b). The key challenge is therefore to translate distribution-level reward extrapolation into the native velocity-regression objective of flow matching.

We address this challenge with **`DreOPD`** (**D**egraded-**r**eference **e**xtrapolative **OPD**), a post-training method that converts implicit reward extrapolation into closed-form velocity regression for flow-matching models. Under shared-covariance Gaussian transitions (Liu et al., 2026), the conditional transition objective at each student-visited state has the pointwise optimizer $v_\lambda^\star = v_T + (\lambda - 1)(v_T - v_{\text{ref}})$. This target recovers teacher imitation at $\lambda = 1$ and moves beyond the teacher when $\lambda > 1$. It further reveals that the teacher-reference contrast determines the direction and magnitude of extrapolation. Motivated by this observation, we construct a mildly degraded reference that enlarges the contrast while preserving the generator's structure. As shown in Figure 1, **`DreOPD`** achieves the best average performance over prior methods and surpasses teachers on most metrics. Our contributions are summarized as follows:

- We derive a closed-form target that extends flow-based OPD from teacher imitation to reward extrapolation.
- We characterize conditions under which the teacher-reference contrast is reward-aligned and introduce controlled reference degradation to strengthen this contrast.
- Across single- and multi-teacher settings, Our **`DreOPD`** achieves the best average performance while surpassing specialized teachers on most metrics.

2 Related Work

2.1 Reinforcement Learning for Flow-matching Models

Reinforcement learning has been increasingly applied to align diffusion and flow-based generative models with task-specific rewards. Early approaches optimize denoising policies through policy gradients or differentiate rewards through the sampling process, including DDPO (Black et al., 2024), DPOK (Fan et al., 2023), DRaFT (Clark et al., 2024), and AlignProp (Prabhudesai et al., 2023). More recent methods, such as Flow-GRPO (Liu et al., 2026) and DanceGRPO (Xue et al., 2025b), adapt group-relative policy optimization (Shao et al., 2024) to stochastic flow trajectories and assign terminal reward signals across continuous denoising steps. These methods are commonly instantiated to optimize a single model against one reward or a fixed aggregation of rewards. Extending them to multiple capabilities requires jointly optimizing heterogeneous objectives, whose gradients may interfere and induce trade-offs among tasks (Li et al., 2025). As a result, RL can be effective for specializing one capability, but multi-task RL remains difficult that the goal is to obtain a single model that performs well across diverse dimensions.

2.2 On-Policy Distillation

On-policy distillation (Agarwal et al., 2024; Li et al., 2026c) trains a student on samples from its current policy, reducing the mismatch between the states encountered during training and generation. In language modeling, methods such as MiniLLM (Gu et al., 2024) and GKD (Tan et al., 2023) use

student-generated sequences together with reverse-KL or generalized divergence objectives to transfer knowledge from a stronger teacher. Vision-OPD (Yuan et al., 2026) extends OPD to multimodal LLMs, using a crop-conditioned teacher to supervise full-image student rollouts for fine-grained visual perception. G-OPD (Yang et al., 2026) interpret on-policy teacher matching as KL-regularized optimization under the implicit reward defined by the teacher-reference log-likelihood ratio, and generalize distillation by scaling this reward beyond exact teacher matching. For flow-matching models, Flow-OPD (Fang et al., 2026) and DiffusionOPD (Li et al., 2026b) instead queries the teacher on states visited by student rollouts and regresses the student toward teacher velocity predictions. DanceOPD (Zhou et al., 2026) further integrates text-to-image and image-to-image capabilities through OPD. Such regression provides dense local supervision, but standard teacher matching makes each teacher prediction the pointwise target (Song & Zheng, 2026). In a multi-task setting, a shared student therefore tends to interpolate among task-specific teachers.

3 Preliminaries

On-policy distillation and reward extrapolation. OPD trains a student distribution π_θ toward a teacher π_T on samples generated by the student itself, typically by minimizing the reverse KL $D_{\mathrm{KL}}(\pi_\theta \| \pi_T)$. G-OPD (Yang et al., 2026) reinterprets teacher matching through the implicit reward $\log(\pi_T/\pi_{\mathrm{ref}})$ and introduces a scaling coefficient λ:

$$\mathbb{E}_{y\sim\pi_\theta}\left[\lambda \log \frac{\pi_T(y)}{\pi_{\mathrm{ref}}(y)} - \log \frac{\pi_\theta(y)}{\pi_{\mathrm{ref}}(y)}\right], \tag{1}$$

where π_{ref} is the reference. At $\lambda = 1$, the reference terms cancel and the objective reduces to reverse-KL teacher matching. The regime $\lambda > 1$, termed ExOPD, amplifies the teacher-reference contrast beyond teacher matching. This formulation establishes a distribution-level principle for reward extrapolation, but does not provide the local velocity target required by flow-matching training.

4 Methodology

Reward extrapolation is naturally expressed for autoregressive language models (Yang et al., 2026), which explicitly parameterize token distributions and provide token-level likelihood ratios. Flow-matching models (Lipman et al., 2022), however, parameterize velocity fields over continuous states, whereas the teacher-reference contrast is defined over complete trajectories. A direct trajectory-level implementation would therefore require likelihood-ratio estimation and long-horizon credit assignment over high-dimensional denoising transitions, making optimization potentially high-variance.

Under shared-covariance Gaussian transitions (Song et al., 2020), we show that the conditional objective at each student-visited state admits the closed-form optimizer,

$$v_\lambda^\star = v_T + (\lambda - 1)(v_T - v_{\mathrm{ref}}), \tag{2}$$

where v_T, v_{ref}, and $v_\lambda^\star$ denote the velocity fields of the teacher, the reference, and the extrapolated target. We train the student by regressing toward the target on its current rollouts. The teacher-reference contrast determines the direction and magnitude of extrapolation, motivating the degraded-reference construction introduced below. Figure 2 provides an overview.

4.1 From Implicit Reward Extrapolation to Flow Velocity Targets

Trajectory-level design objective. Following the reverse-time sampling convention, we discretize the generation schedule as $1 = t_0 > t_1 > \cdots > t_N = 0$ and $\Delta t_j = t_{j+1} - t_j < 0$. A generated trajectory is $\tau = (x_{t_0}, x_{t_1}, \ldots, x_{t_N})$, where x_{t_0} follows the noise prior and x_{t_N} is the generated sample. We distinguish the trajectory distribution Π_θ from its transition kernels $\pi_\theta^{(j)}$:

$$\Pi_\theta(\tau \mid c) = p(x_{t_0}) \prod_{j=0}^{N-1} \pi_\theta^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right), \tag{3}$$

where c denotes the conditioning input. We instantiate the distribution-level principle in Eq. (1) over stochastic flow trajectories, obtaining the design objective

$$\mathcal{J}(\theta) = \lambda\, \mathbb{E}_{\tau\sim\Pi_\theta(\cdot|c)}\left[\log \frac{\Pi_T(\tau \mid c)}{\Pi_{\mathrm{ref}}(\tau \mid c)}\right] - \mathrm{KL}\left(\Pi_\theta(\cdot \mid c) \,\|\, \Pi_{\mathrm{ref}}(\cdot \mid c)\right), \tag{4}$$

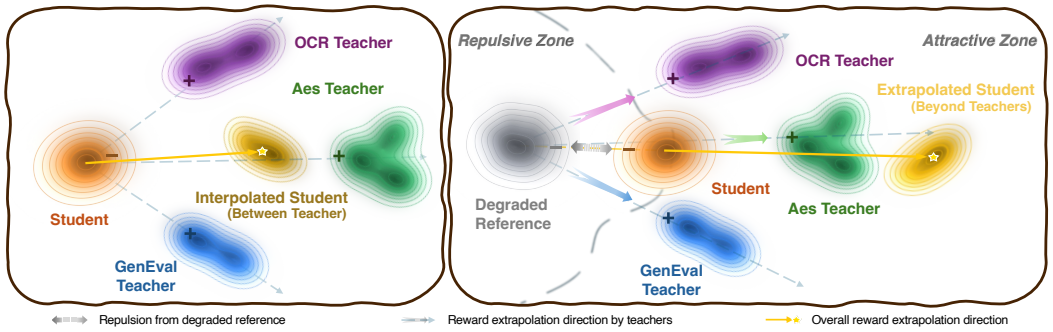


Figure 2: Conceptual comparison in the multi-task setting. Left: Standard OPD regresses a shared student toward task-specific teacher velocities, encouraging interpolation among the teachers; Right: **DreOPD** uses a shared degraded reference to construct targets that extrapolate through each task-specific teacher. Joint regression toward these targets moves the student beyond each teacher along the corresponding teacher-reference directions.

where $\lambda \geq 1$ controls the extrapolation strength. Equivalently,

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \Pi_\theta} \left[\lambda \log \frac{\Pi_T(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} - \log \frac{\Pi_\theta(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} \right]. \tag{5}$$

Directly optimizing Eq. (5) over flow trajectories would require trajectory-level likelihood-ratio estimation and credit assignment across all denoising steps, which can introduce high-variance gradient (Li et al., 2026b). Rather than estimating this trajectory-level policy gradient, we use the shared-covariance Gaussian transition structure to solve the conditional transition objective analytically, obtaining a closed-form velocity target at each student-visited state.

Gaussian flow transitions. Following Flow-GRPO (Liu et al., 2026), the Euler-Maruyama stochastic sampler uses Gaussian reverse-time transitions:

$$\pi_\theta^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right) = \mathcal{N}\left(x_{t_{j+1}}; \mu_\theta^{(j)}, \Sigma_{t_j}\right), \tag{6}$$

with $\mu_\theta = (1 + \frac{\sigma_{t_j}^2 \Delta t_j}{2t_j}) x_{t_j} + (1 + \frac{\sigma_{t_j}^2 (1-t_j)}{2t_j}) v_\theta \Delta t_j$ and $\Sigma_{t_j} = \sigma_{t_j}^2 |\Delta t_j| \, I$. The transition kernels share Σ_{t_j} and differ only in the means induced by their velocity fields.

Let $\mu_\theta^{(j)}$, $\mu_T^{(j)}$, and $\mu_{\text{ref}}^{(j)}$ denote their transition means at a fixed visited state (x_{t_j}, t_j, c). For Gaussian distributions with shared covariance,

$$\begin{aligned} \mathbb{E}_{\pi_\theta^{(j)}}\left[\log \frac{\pi_T^{(j)}}{\pi_{\text{ref}}^{(j)}}\right] &= \frac{\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|^2 - \left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|^2}{2\sigma_{t_j}^2(-\Delta t_j)}, \\ \mathrm{KL}\left(\pi_\theta^{(j)} \,\middle\|\, \pi_{\text{ref}}^{(j)}\right) &= \frac{\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|^2}{2\sigma_{t_j}^2(-\Delta t_j)}. \end{aligned} \tag{7}$$

The common prior $p(x_{t_0})$ cancels from all trajectory likelihood ratios. Using the chain factorization in Eq. (3), the trajectory objective can be written as a sum of conditional transition terms evaluated under student-induced state marginals. For a fixed visited state, maximizing the corresponding objective is equivalent to minimizing

$$\ell_j(\mu_\theta) = \frac{\lambda \left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|^2 - (\lambda - 1)\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|^2}{2\sigma_{t_j}^2(-\Delta t_j)}. \tag{8}$$

Since $\mu_a^{(j)} - \mu_b^{(j)} = \Delta t_j \, (v_a - v_b)$, the same conditional objective can be expressed in velocity space with a weight κ_{t_j}:

$$\ell_j(v_\theta) = \kappa_{t_j}[\lambda \left\|v_\theta - v_T\right\|^2 - (\lambda - 1)\left\|v_\theta - v_{\text{ref}}\right\|^2], \quad \kappa_{t_j} = -\Delta t_j / 2\sigma_{t_j}^2 > 0. \tag{9}$$

Although Eq. (9) contains a negative quadratic term, its net coefficient on $\|v_\theta\|^2$ is $\lambda - (\lambda - 1) = 1$. Therefore, the objective remains strongly convex in v_θ and admits a unique finite minimizer. Completing the square gives

$$\ell_j(v_\theta) = \kappa_{t_j} \|v_\theta - v_\lambda^\star\|^2 + C_j, \tag{10}$$

where C_j does not depend on v_θ and

$$v_\lambda^\star = v_T + (\lambda - 1)(v_T - v_{\text{ref}}). \tag{11}$$

Equation (11) is the central target of **DreOPD**. At $\lambda = 1$, it reduces to $v_\lambda^\star = v_T$, recovering teacher imitation. For $\lambda > 1$, the target moves beyond the teacher along the teacher-reference direction. The displacement from the teacher is

$$v_\lambda^\star - v_T = (\lambda - 1)(v_T - v_{\text{ref}}), \tag{12}$$

which makes the influence of both the extrapolation strength and the reference explicit.

On-policy regression objective. Given the extrapolated velocity target $v_\lambda^\star$, we optimize the target by on-policy regression objective:

$$\mathcal{L}(\theta) = \mathbb{E}_c\left[\sum_{j=0}^{N-1} \kappa_{t_j}\left(\left\|v_\theta(x_{t_j}, t_j, c) - \text{sg}(v_\lambda^\star(x_{t_j}, t_j, c))\right\|^2\right)\right], \tag{13}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The trajectory $\{x_{t_j}\}_{j=0}^N$ is generated by the current student rollouts. Furthermore, under the deterministic ODE, the student, teacher and reference induce deterministic velocity predictions. We therefore optimize the same closed-form target using direct ℓ_2 regression, weighted by $(\Delta t_j)^2/2$. Appendix B.2 details the connection.

4.2 Reward Interpretation of the Extrapolated Target

The preceding derivation establishes a geometric result: the proposed target extrapolates from the reference through the teacher. This geometric property alone does not guarantee improvement under an arbitrary task reward. We therefore characterize a sufficient condition under which the teacher-reference contrast is reward-aligned. The analysis is stated at the terminal-distribution level and provides a reward interpretation for the contrast locally amplified by Eq. (11).

Lemma 1 (Teacher as a reward-tilted reference)**.** *Let $r(x)$ be a bounded reward and $\beta > 0$ a temperature. Suppose the teacher solves the KL-regularized reward optimization problem $p_T = \arg\max_p \{\mathbb{E}_{x\sim p}[r(x)] - \beta\,\text{KL}(p \,\|\, p_{\text{ref}})\}$. Then $p_T(x) = (1/Z_T)\; p_{\text{ref}}(x)\exp(r(x)/\beta)$, where $Z_T = \int p_{\text{ref}}(x)\exp(r(x)/\beta)\,dx$ is a normalizing constant.*

Lemma 1 implies $\log[p_T(x)/p_{\text{ref}}(x)] = r(x)/\beta - \log Z_T$. Under this assumption, the teacher-reference log-density ratio recovers the reward $r(x)$ up to a positive scale and an additive constant.

Proposition 2 (Distributional optimum and reward monotonicity)**.** *Consider the distribution-level objective*

$$\mathcal{J}_\lambda(p) = \lambda\,\mathbb{E}_{x\sim p}\left[\log\frac{p_T(x)}{p_{\text{ref}}(x)}\right] - \text{KL}(p \,\|\, p_{\text{ref}}), \tag{14}$$

where $\lambda \geq 1$. Assuming the normalizing constant is finite, its optimizer over distributions absolutely continuous with respect to p_{ref} is $p_\lambda(x) = \frac{1}{Z_\lambda} p_T(x)^\lambda p_{\text{ref}}(x)^{1-\lambda}$. Moreover, under the reward-tilt assumption of Lemma 1, the expected reward $J(\lambda) = \mathbb{E}_{x\sim p_\lambda}[r(x)]$ satisfies

$$\frac{dJ(\lambda)}{d\lambda} = \frac{1}{\beta}\text{Var}_{x\sim p_\lambda}[r(x)] \geq 0, \tag{15}$$

with strict inequality whenever r is non-constant p_λ-almost surely.

Proofs are provided in Appendix B.3 and Appendix B.4. The optimal distribution satisfies $\log p_\lambda = \log p_T + (\lambda - 1)(\log p_T - \log p_{\text{ref}}) - \log Z_\lambda$. Thus, $\lambda > 1$ extrapolates beyond the teacher along the teacher-reference log-density contrast. Under the reward-tilt assumption, this contrast is proportional to the underlying reward up to scale and a constant, yielding the monotonicity result in Eq. (15). Proposition 2 establishes the reward interpretation of extrapolation at the distribution level, and our Gaussian transition derivation translates this principle into the extrapolative velocity target used by our **DreOPD**.

4.3 Amplifying Extrapolation with a Degraded Reference

Eq. (12) shows that the displacement beyond the teacher scales with $v_T - v_{\text{ref}}$. Reference selection therefore determines the direction and magnitude of the extrapolation.

Reference determines the extrapolation contrast. In standard KL-regularized optimization, the reference acts primarily as a conservative anchor that discourages large policy changes. In Eq. (11), it has an additional geometric role. The teacher attracts the student, while the reference specifies the direction from which the target moves through and beyond the teacher. A reference close to the teacher yields a small contrast, making the extrapolation signal weak. This motivates using a mildly degraded reference to enlarge teacher-reference contrast while retains reward-aligned directions.

Reward-alignment model. We formalize this intuition through an idealized family of reward-aligned distributions. Let p_0 denote the distribution induced by the pretrained generator that initializes the student, and define

$$p_\rho(x) = \frac{1}{Z_\rho} p_0(x) \exp\left(\frac{\rho\, r(x)}{\beta}\right), \qquad \rho \le 1, \tag{16}$$

where ρ denotes the level of alignment with reward r. We parameterize the teacher as $p_T = p_1$ and the pretrained model as p_0. A degraded reference with lower reward alignment than the pretrained model is represented by $\rho < 0$.

Proposition 3 (Reference degradation amplifies extrapolation)**.** *Let $p_T = p_1$ and $p_{\text{ref}} = p_\rho$ be members of the family in Eq. (16). Then the extrapolated distribution $p_{\lambda,\rho}(x) \propto p_T(x)^\lambda p_\rho(x)^{1-\lambda}$ is also a reward tilt of p_0:*

$$p_{\lambda,\rho}(x) = \frac{1}{Z_{\lambda,\rho}} p_0(x) \exp\left(\frac{\rho_{\text{eff}}(\lambda,\rho) r(x)}{\beta}\right), \tag{17}$$

where $\rho_{\text{eff}}(\lambda,\rho) = 1+(\lambda-1)(1-\rho)$. Therefore, for a degraded reference ρ_d and a normal reference ρ_n satisfying $\rho_d < \rho_n < 1$, any fixed $\lambda > 1$ yields $\rho_{\text{eff}}(\lambda,\rho_d) > \rho_{\text{eff}}(\lambda,\rho_n) > 1$. Moreover, if r is nonconstant under p_T, their local reward sensitivities at $\lambda = 1$ satisfy

$$\left.\frac{dJ_{\rho_d}(\lambda)/d\lambda}{dJ_{\rho_n}(\lambda)/d\lambda}\right|_{\lambda=1} = \frac{1-\rho_d}{1-\rho_n} > 1, \; J_\rho(\lambda) = \mathbb{E}_{p_{\lambda,\rho}}[r(x)]. \tag{18}$$

Proofs are in Appendix B.5. At $\lambda = 1$, every reference choice recovers the teacher distribution, whereas the derivative with respect to λ scales with the contrast magnitude $1 - \rho$. Proposition 3 therefore shows that, within the reward-alignment model, a lower-alignment reference increases both the effective reward-tilt coefficient and the local rate of reward change without altering the reward direction. This result motivates controlled reference degradation for strengthening extrapolation.

Practical degraded-reference construction. In practice, we construct the degraded reference from the same base model using controlled corruption such as weight quantization or mild noise injection into the velocity output. The degradation level must balance contrast strength against structural preservation. Insufficient degradation produces a weak teacher-reference contrast, whereas excessive or poorly structured degradation may cause $v_T - v_{\text{ref}}$ to capture corruption artifacts rather than task-relevant differences. We therefore evaluate multiple degradation mechanisms.

4.4 Training procedure

Given task-specific teachers trained by existing RL algorithms (Zheng et al., 2025b; Wang et al., 2026) and a degraded reference, each training round iterates over the tasks. For each task, the current student generates on-policy trajectories and the corresponding teacher and reference are queried at the visited states. Finally, the student is updated toward the closed-form target. Algorithm 1 in the Appendix C summarizes the training procedure of **`DreOPD`**.

Table 1: Single-teacher distillation on three tasks. GenEval and OCR are reported by their respective students, while PickScore, ClipScore, and HPSv2.1 are reported by the aesthetics student. The **`Teacher`** row similarly combines the in-domain scores of three task-specific teachers. **Bold**: best; Underline: second best. Subscripts show relative changes changes from the corresponding teacher.

Method	GenEval Student	OCR Student	Aesthetic Student		
	GenEval	OCR	PickScore	ClipScore	HPSv2.1
`SD3.5-M (w/o CFG)`	0.2529	0.1377	20.519	0.2384	0.2052
`SD3.5-M`	0.6273	0.5079	22.331	0.2837	0.2795
`Teacher`	0.9470	0.9239	24.034	**0.2963**	0.3460
`Flow-OPD`	$0.9501_{+0.33\%}$	$0.9279_{+0.43\%}$	$24.006_{-0.12\%}$	$0.2957_{-0.20\%}$	$0.3445_{-0.43\%}$
`DiffusionOPD`	$0.9648_{+1.88\%}$	$0.9246_{+0.08\%}$	$24.008_{-0.11\%}$	$0.2955_{-0.27\%}$	$0.3454_{-0.17\%}$
`Ours`	$\mathbf{0.9710}_{+2.53\%}$	$\mathbf{0.9364}_{+1.35\%}$	$\mathbf{24.037}_{+0.01\%}$	$0.2955_{-0.27\%}$	$\mathbf{0.3497}_{+1.07\%}$

5 Experiments

5.1 Experimental Setup

Implementation details. We conduct all experiments with SD3.5-M (Esser et al., 2024) at a resolution of 512×512, keeping the teacher and reference models frozen throughout distillation. GenEval (Ghosh et al., 2023) and OCR use the data splits released with FlowGRPO (Liu et al., 2026), while aesthetic optimization is performed on Pick-a-Pic (Kirstain et al., 2023) and evaluated on DrawBench (Saharia et al., 2022). We also report Aesthetic (Schuhmann, 2022) and ImageReward (Xu et al., 2023) as out-of-domain reward. For comparability with prior work, we follow the configuration of DiffusionOPD, with additional details provided in Appendix D.

Single-task RL teachers. We use the same specialized teachers as DiffusionOPD (Li et al., 2026b). The GenEval teacher is trained with DiffusionNFT (Zheng et al., 2025b), while the OCR and aesthetics teachers are trained with GRPO-Guard (Wang et al., 2026). The aesthetics teacher jointly optimizes an equally weighted combination of PickScore, ClipScore (Hessel et al., 2021), and HPSv2.1 (Wu et al., 2023).

Multi-task RL baselines. We compare against Flow-GRPO, GRPO-Guard, and DiffusionNFT trained directly in the multi-task setting. These methods optimize a single model over the three tasks by alternating among their corresponding datasets. We additionally include CascadeNFT, which optimizes the tasks sequentially.

OPD baselines. For a fair comparison, all OPD methods use the same task-specific teachers. Flow-OPD (Fang et al., 2026) performs teacher-oriented velocity regression and uses Manifold Anchor Regularization to constrain the student to a high-quality visual manifold. DiffusionOPD (Li et al., 2026b) derives a closed-form reverse-KL objective for consolidating specialized teachers.

5.2 Single-Teacher Distillation

We investigate whether on-policy distillation can surpass specialized teachers within their respective domains. Table 1 summarizes three independent experiments: separate students are distilled for GenEval, OCR, and aesthetics, with the aesthetics student evaluated by PickScore, ClipScore, and HPSv2.1. Accordingly, the **`Teacher`** row combines the in-domain scores of the three corresponding task-specific teachers rather than representing a single teacher model.

The conventional OPD baselines remain close to their corresponding teachers. Flow-OPD yields small gains on GenEval and OCR but slightly decreases all three aesthetics metrics. DiffusionOPD improves GenEval more substantially, while remaining close to the OCR teacher and slightly underperforming the aesthetics teacher on its three evaluation metrics. These results are consistent with objectives centered primarily on teacher matching. In contrast, our method improves GenEval from 0.9470 to 0.9710, OCR from 0.9239 to 0.9364, PickScore from 24.034 to 24.037 and HPSv2.1 from 0.3460 to 0.3497. Overall, it surpasses the corresponding teachers on four of five metrics.

Table 2: Results of single-task RL, multi-task RL and OPD methods. **SD3.5-M**†: without CFG; **Bold**: best; Underline: second-best; Gray-colored: In-Domain reward. Subscripts show relative changes from the corresponding teacher. Out-Of-Domain rewards: Aesthetic and ImgReward.

Model	GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImgReward
SD3.5-M†	0.2529	0.1377	20.519	0.2384	0.2052	5.161	-0.5471
SD3.5-M	0.6273	0.5079	22.331	0.2837	0.2795	5.396	0.8324
Single-Task RL Teachers							
GenEval	0.9470	0.6286	20.084	0.2870	0.2644	5.246	0.8976
OCR	0.6562	0.9239	22.225	0.2919	0.2720	5.266	0.8881
Aes	0.4935	0.5014	24.034	0.2963	0.3460	6.232	1.5071
Multi-Task RL							
Flow-GRPO	0.7399	0.7673	22.677	0.2893	0.2901	5.469	1.0539
GRPO-Guard	0.9002	0.9278	23.197	0.2959	0.3137	5.824	1.3582
DiffNFT	0.9510	**0.9491**	23.182	0.2869	0.2812	5.399	1.0699
CascadeNFT	0.9376	0.8827	23.803	0.2920	0.3305	6.006	1.4907
On-Policy Distillation							
Flow-OPD	0.9395$_{-0.79\%}$	0.8756$_{-5.23\%}$	23.262$_{-3.21\%}$	**0.2975**$_{+0.41\%}$	0.3070$_{-11.3\%}$	5.692$_{-8.67\%}$	1.3251$_{-12.8\%}$
DiffOPD	0.9607$_{+1.45\%}$	0.9242$_{+0.03\%}$	23.980$_{-0.23\%}$	0.2962$_{-0.03\%}$	0.3422$_{-1.10\%}$	6.191$_{-0.66\%}$	1.5017$_{-0.36\%}$
Ours	**0.9681**$_{+2.23\%}$	0.9362$_{+1.33\%}$	**24.035**$_{+0.01\%}$	0.2959$_{-0.14\%}$	**0.3487**$_{+0.78\%}$	**6.292**$_{+0.96\%}$	**1.5245**$_{+1.16\%}$



Figure 3: Qualitative comparison of the base model, single-task teachers, multi-task RL, OPD baselines, and our method.

5.3 Multi-Teacher Distillation

Table 2 compares task-specific teachers, multi-task RL, and multi-teacher OPD. Each teacher excels in its own domain but transfers poorly to other objectives. Each teacher excels in its own domain but transfers poorly across objectives, motivating their consolidation into a single generalist model.

Multi-task RL improves task coverage but suffers from optimization interference. DiffusionNFT achieves the highest OCR score at the expense of perceptual quality, while GRPO-Guard is more balanced but remains suboptimal overall. CascadeNFT performs better through sequential optimization but still falls short of the strongest distillation methods. These results reveal the task conflict induced by jointly optimizing heterogeneous rewards.

Multi-teacher OPD instead separates teacher specialization from knowledge consolidation. Flow-OPD underperforms the teachers on several metrics, while DiffusionOPD largely preserves their capabilities. Our **DreOPD** achieves the best overall performance, outperforming the corresponding teachers on GenEval, OCR, HPSv2.1, Aesthetic and ImageReward metrics while matching PickScore and retaining comparable ClipScore. Figure 3 shows qualitative results.

5.4 Ablation Studies

Noise level. As shown in Figure 4, reducing noise level consistently improves performance. Deterministic ODE sampling performs best, followed by noise levels of 0.3, 0.5 and 0.7. This suggests that the closed-form target does not require stochastic exploration to provide an effective learning signal. At the same noise level of 0.3, the policy-gradient baseline performs the worst, supporting our analysis that direct policy optimization introduces additional variance and lead to inferior results.

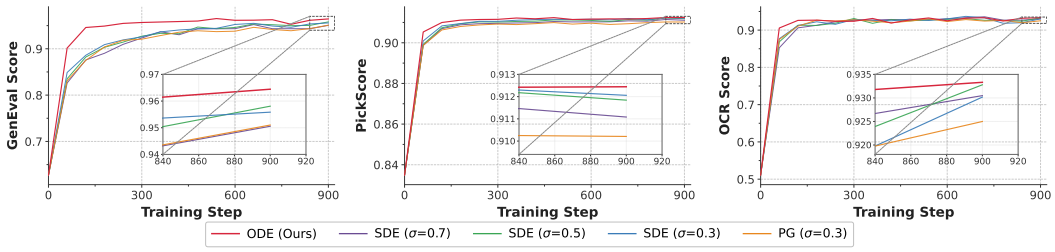


Figure 4: Ablation results on the noise level. We compare deterministic ODE sampling with SDE sampling at noise levels 0.3, 0.5, and 0.7, as well as a policy-gradient variant at noise level 0.3, on GenEval, PickScore, and OCR.

Table 3: Ablation results on degraded reference. Rewards: GenEval, OCR, PickScore, ClipScore, HPSv2.1, Aesthetic and **ImageReward**. **Bold**: best; Underline: second best.

Method	In-Domain					OOD		Avg.
	GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImgRwd	
w/o DeRef.	0.967	0.928	24.03	0.2956	0.3487	6.24	1.517	0.558
Reference Velocity Quantization								
4-bit	0.962	0.923	24.03	0.2950	0.3482	6.28	1.524	0.566
8-bit	**0.968**	**0.936**	24.04	**0.2959**	0.3487	6.29	**1.525**	**0.926**
Reference Weight Quantization								
4-bit	0.959	0.920	23.96	0.2929	0.3489	**6.30**	1.523	0.372
8-bit	0.961	0.931	**24.05**	0.2949	0.3490	6.28	**1.525**	0.739
Reference Velocity Gaussian Perturbation								
$\sigma = 0.01$	0.959	0.923	24.01	0.2954	0.3477	6.27	1.520	0.358
$\sigma = 0.1$	0.960	0.925	24.04	0.2948	**0.3491**	6.28	1.521	0.594

Degraded reference. Table 3 compares different degradation strategies. Moderate degradation performs best, with 8-bit velocity quantization achieving the highest average score. This suggests that a mildly degraded reference can enlarge the teacher-reference contrast and provide a more informative extrapolation direction. However, overly weak degradation gives limited contrast, whereas stronger degradation disrupts the reference structure and yields inconsistent gains by making the extrapolation direction less reliable. Appendix D.4 reports the metrics of each degraded reference.

Impact of factor λ. Table 4 studies the effect of λ using the original reference. Interpolation ($\lambda < 1$) generally underperforms teacher matching ($\lambda = 1$), whereas extrapolation ($\lambda > 1$) performs better overall. We use $\lambda = 1.25$, which provides the best cross-task balance. Although $\lambda = 1.5$ further improves OCR, it degrades GenEval and perceptual quality, suggesting that excessive extrapolation may overemphasize one capability at the cost of overall alignment.

Table 4: Ablation on factor λ. Rewards: **G**en**E**val, **OCR**, **Pick**Score, **Clip**Score, **HPS**v2.1, **Aes**thetic and **I**mage**R**eward.

λ	In-Domain					OOD	
	GE	OCR	Pick	Clip	HPS	Aes	IR
0.5	0.921	0.899	23.707	0.2941	0.3241	6.03	1.411
0.75	0.950	0.918	23.925	0.2961	0.3364	6.13	1.483
1.0	0.958	0.926	23.988	**0.2963**	0.3426	6.20	1.504
1.25	**0.968**	0.936	**24.034**	0.2959	0.3487	**6.29**	**1.525**
1.5	0.961	**0.938**	24.019	0.2945	**0.3505**	6.26	1.523

6 Conclusion

We introduced **DreOPD**, a degraded-reference extrapolative on-policy distillation method for flow-matching models. **DreOPD** translates trajectory-level implicit reward extrapolation into a closed-form velocity target, extending OPD from teacher imitation to teacher-reference extrapolation while retaining regression-based training. We further showed that a mildly degraded reference can strengthen a reward-aligned contrast without disrupting generative structure. Across multiple settings, **DreOPD** achieves the best average performance over the baselines, while surpassing teachers on most metrics. These results establish reward extrapolation as an effective framework for consolidating specialized flow models while improving beyond teacher imitation.

AI USE STATEMENT

The authors determined the research questions, methodology, experimental design, analysis, and claims of this work. Generative AI tools were used to assist with identifying potentially relevant literature and with manuscript preparation, including grammar checking, sentence-level editing, and polishing for clarity and readability. The authors independently verified all cited sources and take full responsibility for the final content of the paper, including its text, experiments, results, and conclusions.

ETHIC STATEMENT

Our work improves the post-training of text-to-image generative models and may inherit biases, harmful associations, or misuse risks from the underlying pretrained models and reward functions. We conduct experiments using established models and benchmarks and do not collect personal or sensitive data. While our method is intended for research on model alignment and capability consolidation, it could potentially enhance the generation of misleading or harmful content. We encourage responsible deployment with appropriate safety evaluation, content safeguards, and consideration of representational biases.

REPRODUCIBILITY STATEMENT

We provide details of the experimental setup, model configurations, evaluation metrics, and implementation procedures in the paper and appendix. We will release code and configuration files to facilitate reproduction of the main experiments and analyses.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Shengchao Hu, Ziqing Fan, Li Shen, Ya Zhang, Yanfeng Wang, and Dacheng Tao. Harmodt: harmony multi-task decision transformer for offline reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19182–19197, 2024.

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. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Chaoyu Li, Xiaoyi Gu, Yogesh Kulkarni, Eun Woo Im, Mohammadmahdi Honarmand, Zeyu Wang, Juntong Song, Fei Du, Xilin Jiang, Kexin Zheng, et al. Video generation models: A survey of post-training and alignment. *Transactions on Machine Learning Research*, 2026a.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

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*, 2026b.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026c.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

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. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Lei Lyu, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. *Advances in Neural Information Processing Systems*, 38:93967–93990, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann. Laion-aesthetics, 2022. URL https://laion.ai/blog/laion-aesthetics/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, and Jie Tang. Gkd: A general knowledge distillation framework for large-scale pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 134–148, 2023.

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.

Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, et al. Grpo-guard: Mitigating implicit over-optimization in flow matching via regulated clipping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5988–5998, 2026.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025a.

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

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Model extrapolation expedites alignment. In *63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, pp. 1025–1041. Association for Computational Linguistics (ACL), 2025a.

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*, 2025b.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, et al. Danceopd: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026.

A TEACHER SURPASSING VISUALIZATION

To further visualize the multi-task capability integration, we provide a radar plot in Figure 5. Our **DreOPD** outperforms the corresponding teacher on six out of seven metrics, with only a slight decrease on ClipScore, and achieves the highest average score overall. This indicates that our student does not merely average teacher behaviors, but successfully consolidates and extrapolates their strengths across tasks.

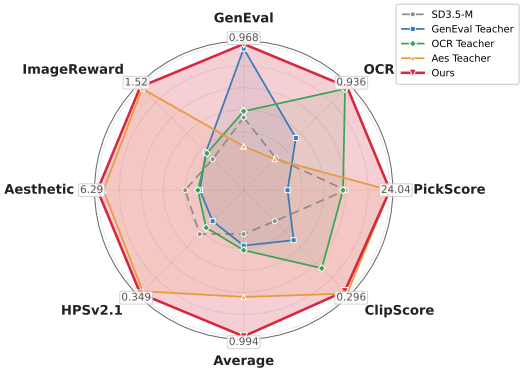


Figure 5: Visualization of multi-task **DreOPD** performance. We compare the student trained by **DreOPD** with the corresponding task-specific teachers across metrics.

B DERIVATIONS AND PROOFS

B.1 DERIVATION OF THE CLOSED-FORM VELOCITY TARGET

This section provides the full derivation of the closed-form velocity target in Eq. (11).

$$v_\lambda^\star = v_T + (\lambda - 1)\,(v_T - v_{\text{ref}})\,.$$

We first decompose the trajectory-level objective into conditional transition objectives and then solve each conditional objective under the shared-covariance Gaussian transition model. Finally, we clarify the relationship between the resulting pointwise optimizer and the on-policy regression objective used for training.

B.1.1 TRAJECTORY-LEVEL OBJECTIVE DECOMPOSITION

Fix a conditioning input c. Following the reverse-time sampling convention, let

$$1 = t_0 > t_1 > \cdots > t_N = 0, \qquad \Delta t_j = t_{j+1} - t_j < 0.$$

A trajectory is denoted by $\tau = (x_{t_0}, x_{t_1}, \ldots, x_{t_N})$. The student, teacher, and reference trajectory distributions factorize as

$$\Pi_a(\tau \mid c) = p(x_{t_0}) \prod_{j=0}^{N-1} \pi_a^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right), \tag{19}$$

where $a \in \{\theta, T, \text{ref}\}$ and all three models share the same initial noise distribution $p(x_{t_0})$.

The trajectory-level extrapolation objective is

$$\mathcal{J}(\theta; c) = \mathbb{E}_{\tau \sim \Pi_\theta(\cdot|c)} \left[\lambda \log \frac{\Pi_T(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} - \log \frac{\Pi_\theta(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} \right]. \tag{20}$$

Because the initial distribution $p(x_{t_0})$ is shared, it cancels from each trajectory likelihood ratio. In particular,

$$\log \frac{\Pi_T(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} = \log \frac{p(x_{t_0}) \prod_{j=0}^{N-1} \pi_T^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)}{p(x_{t_0}) \prod_{j=0}^{N-1} \pi_{\text{ref}}^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)} = \sum_{j=0}^{N-1} \log \frac{\pi_T^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)}{\pi_{\text{ref}}^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)},$$
$$\log \frac{\Pi_\theta(\tau \mid c)}{\Pi_{\text{ref}}(\tau \mid c)} = \sum_{j=0}^{N-1} \log \frac{\pi_\theta^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)}{\pi_{\text{ref}}^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right)}. \tag{21}$$

Let $d_{\theta,j}(x_{t_j} \mid c)$ denote the marginal distribution of x_{t_j} induced by the student trajectory distribution $\Pi_\theta(\cdot \mid c)$. Substituting Eq. (21) into Eq. (20) and applying the tower property gives

$$\mathcal{J}(\theta) = \sum_{j=0}^{N-1} \mathbb{E}_{x_{t_j} \sim d_{\theta,j}(\cdot|c)} \left[\lambda \, \mathbb{E}_{x_{t_{j+1}} \sim \pi_\theta^{(j)}} \left[\log \frac{\pi_T^{(j)}}{\pi_{\text{ref}}^{(j)}} \right] - \mathrm{KL}\left(\pi_\theta^{(j)}(\cdot \mid x_{t_j}, c) \,\middle\|\, \pi_{\text{ref}}^{(j)}(\cdot \mid x_{t_j}, c) \right) \right]. \tag{22}$$

Equation (22) is an exact decomposition of the trajectory-level objective.

B.1.2 Conditional Gaussian Transition Objective

At a fixed state (x_{t_j}, t_j, c), suppose that the student, teacher, and reference transitions are Gaussian with a shared covariance:

$$\pi_a^{(j)}\left(x_{t_{j+1}} \mid x_{t_j}, c\right) = \mathcal{N}\left(x_{t_{j+1}}; \mu_a^{(j)}, \Sigma_{t_j}\right),$$
$$\mu_a^{(j)} = x_{t_j} + v_a(x_{t_j}, t_j, c)\Delta t_j, \Sigma_{t_j} = \sigma_{t_j}^2 |\Delta t_j| I = \sigma_{t_j}^2 (-\Delta t_j) I. \tag{23}$$

The final equality follows from $\Delta t_j < 0$. For Gaussian distributions with the same covariance matrix,

$$\mathrm{KL}\left(\mathcal{N}(\mu_a, \Sigma) \,\|\, \mathcal{N}(\mu_b, \Sigma)\right) = \frac{1}{2}(\mu_a - \mu_b)^\top \Sigma^{-1}(\mu_a - \mu_b). \tag{24}$$

Moreover, for $X \sim \mathcal{N}(\mu_\theta, \Sigma)$,

$$\begin{aligned}
&\mathbb{E}\left[\log \frac{\mathcal{N}(X; \mu_T, \Sigma)}{\mathcal{N}(X; \mu_{\text{ref}}, \Sigma)}\right] \\
&= -\frac{1}{2}\mathbb{E}\left[(X - \mu_T)^\top \Sigma^{-1}(X - \mu_T)\right] + \frac{1}{2}\mathbb{E}\left[(X - \mu_{\text{ref}})^\top \Sigma^{-1}(X - \mu_{\text{ref}})\right] \\
&= -\frac{1}{2}\Big[\operatorname{tr}(\Sigma^{-1}\Sigma) + (\mu_\theta - \mu_T)^\top \Sigma^{-1}(\mu_\theta - \mu_T)\Big] + \frac{1}{2}\Big[\operatorname{tr}(\Sigma^{-1}\Sigma) + (\mu_\theta - \mu_{\text{ref}})^\top \Sigma^{-1}(\mu_\theta - \mu_{\text{ref}})\Big] \\
&= \frac{1}{2}\|\mu_\theta - \mu_{\text{ref}}\|_{\Sigma^{-1}}^2 - \frac{1}{2}\|\mu_\theta - \mu_T\|_{\Sigma^{-1}}^2,
\end{aligned} \tag{25}$$

where $\|z\|_{\Sigma^{-1}}^2 = z^\top \Sigma^{-1} z$. The trace terms cancel because all three transitions share the same covariance.

Applying Eq. (24) and Eq. (25) at step j, the term of expectation in Eq. (22) becomes

$$\begin{aligned}
\mathcal{J}_j^{\text{cond}}(\mu_\theta^{(j)}) &= \lambda \, \mathbb{E}_{\pi_\theta^{(j)}}\left[\log \frac{\pi_T^{(j)}}{\pi_{\text{ref}}^{(j)}}\right] - \mathrm{KL}\left(\pi_\theta^{(j)} \,\middle\|\, \pi_{\text{ref}}^{(j)}\right) \\
&= \underbrace{\frac{\lambda}{2}\left(\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2 - \left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2\right)}_{\lambda \, \mathbb{E}\left[\log \pi_T^{(j)}/\pi_{\text{ref}}^{(j)}\right]} - \underbrace{\frac{1}{2}\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2}_{\mathrm{KL}\left(\pi_\theta^{(j)} \,\|\, \pi_{\text{ref}}^{(j)}\right)} \\
&= -\frac{1}{2}\left[\lambda \left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2 - (\lambda - 1)\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2\right].
\end{aligned} \tag{26}$$

Therefore, under $\Sigma_{t_j}^{-1} = [\sigma_{t_j}^2(-\Delta t_j)]^{-1} I$, maximizing the conditional objective is equivalent to minimizing

$$\begin{aligned}\ell_j(\mu_\theta) &= \frac{1}{2}\left[\lambda \left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2 - (\lambda-1)\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|_{\Sigma_{t_j}^{-1}}^2\right] \\ &= \frac{\lambda\left\|\mu_\theta^{(j)} - \mu_T^{(j)}\right\|^2 - (\lambda-1)\left\|\mu_\theta^{(j)} - \mu_{\text{ref}}^{(j)}\right\|^2}{2\sigma_{t_j}^2(-\Delta t_j)}.\end{aligned} \tag{27}$$

B.1.3 Closed-Form Optimizer in Velocity Space

From Eq. (23), the difference between any two transition means satisfies

$$\mu_a^{(j)} - \mu_b^{(j)} = \Delta t_j \left[v_a(x_{t_j}, t_j, c) - v_b(x_{t_j}, t_j, c)\right]. \tag{28}$$

Substituting Eq. (28) into Eq. (27) gives

$$\begin{aligned}\ell_j(v_\theta) &= \frac{(\Delta t_j)^2}{2\sigma_{t_j}^2(-\Delta t_j)}\left[\lambda\left\|v_\theta - v_T\right\|^2 - (\lambda-1)\left\|v_\theta - v_{\text{ref}}\right\|^2\right] \\ &= \kappa_{t_j}\left[\lambda\left\|v_\theta - v_T\right\|^2 - (\lambda-1)\left\|v_\theta - v_{\text{ref}}\right\|^2\right],\end{aligned} \tag{29}$$

where $\kappa_{t_j} = -\Delta t_j / 2\sigma_{t_j}^2 > 0$. Although Eq. (29) contains a negative quadratic term, the overall objective is strictly convex in v_θ. To see this, expand the two squared distances:

$$\begin{aligned}\frac{\ell_j(v_\theta)}{\kappa_{t_j}} &= \lambda\left(\|v_\theta\|^2 - 2v_\theta^\top v_T + \|v_T\|^2\right) - (\lambda-1)\left(\|v_\theta\|^2 - 2v_\theta^\top v_{\text{ref}} + \|v_{\text{ref}}\|^2\right) \\ &= \|v_\theta\|^2 - 2v_\theta^\top\left[\lambda v_T - (\lambda-1)v_{\text{ref}}\right] + \lambda\|v_T\|^2 - (\lambda-1)\|v_{\text{ref}}\|^2.\end{aligned} \tag{30}$$

The coefficient of $\|v_\theta\|^2$ is $\lambda - (\lambda-1) = 1$. Consequently, the Hessian is $\nabla_{v_\theta}^2 \ell_j(v_\theta) = 2\kappa_{t_j} I \succ 0$ and the conditional objective has a unique finite minimizer.

Define $v_\lambda^\star = \lambda v_T - (\lambda-1)v_{\text{ref}} = v_T + (\lambda-1)(v_T - v_{\text{ref}})$. Completing the square in Eq. (30) yields

$$\begin{aligned}\frac{\ell_j(v_\theta)}{\kappa_{t_j}} &= \|v_\theta - v_\lambda^\star\|^2 + \lambda\|v_T\|^2 - (\lambda-1)\|v_{\text{ref}}\|^2 - \|v_\lambda^\star\|^2 \\ &= \|v_\theta - v_\lambda^\star\|^2 - \lambda(\lambda-1)\left\|v_T - v_{\text{ref}}\right\|^2.\end{aligned} \tag{31}$$

The second term $\lambda(\lambda-1)\left\|v_T - v_{\text{ref}}\right\|^2$ is independent of v_θ. Hence, $\arg\min_{v_\theta} \ell_j(v_\theta) = v_\lambda^\star$.

B.2 From SDE to ODE

Our closed-form target is derived from stochastic transitions with a shared Gaussian covariance. In practice, we follow DiffusionOPD (Li et al., 2026b) and use deterministic ODE rollouts, which perform best in our ablation. The SDE conditional objective can be written as

$$\ell_j^{\text{SDE}}(v_\theta) = \kappa_{t_j}\left[\lambda\|v_\theta - v_T\|^2 - (\lambda-1)\|v_\theta - v_{\text{ref}}\|^2\right] = \kappa_{t_j}\left\|v_\theta - v_\lambda^\star\right\|^2 + C_j, \tag{32}$$

where $v_\lambda^\star = v_T + (\lambda-1)(v_T - v_{\text{ref}})$, $\kappa_{t_j} = -\Delta t_j/(2\sigma_{t_j}^2) > 0$ and C_j is independent of v_θ at a fixed visited state. Therefore, the pointwise optimizer $v_\lambda^\star$ is independent of the shared transition noise.

Under the deterministic ODE update, the student, teacher, and reference induce the transition means

$$\mu_a^{(j)} = x_{t_j} + v_a(x_{t_j}, t_j, c)\Delta t_j, \quad a \in \{\theta, T, \text{ref}\}. \tag{33}$$

The extrapolated transition target is therefore

$$\mu_\lambda^\star = \mu_T^{(j)} + (\lambda-1)\left(\mu_T^{(j)} - \mu_{\text{ref}}^{(j)}\right) = x_{t_j} + v_\lambda^\star(x_{t_j}, t_j, c)\Delta t_j. \tag{34}$$

For the deterministic transition-matching objective, we directly regress the student transition mean toward $\mu_\lambda^\star$. Since $\mu_\theta^{(j)} - \mu_\lambda^\star = \Delta t_j (v_\theta - v_\lambda^\star)$, the ODE objective becomes

$$\begin{aligned}\mathcal{L}^{\mathrm{ODE}}(\theta) &= \mathbb{E}_c \left[\sum_{j=0}^{N-1} \frac{1}{2} \left\| \mu_\theta^{(j)} - \mathrm{sg}(\mu_\lambda^\star) \right\|^2 \right] \\ &= \mathbb{E}_c \left[\sum_{j=0}^{N-1} \frac{(\Delta t_j)^2}{2} \left\| v_\theta(x_{t_j}, t_j, c) - \mathrm{sg}\big(v_\lambda^\star(x_{t_j}, t_j, c)\big) \right\|^2 \right].\end{aligned} \tag{35}$$

The states $\{x_{t_j}\}_{j=0}^N$ are collected from the current student ODE rollout and treated as fixed during each regression update. Thus, the ODE formulation retains the closed-form extrapolated target while replacing stochastic transition matching with direct deterministic mean matching.

B.3 Proof of Lemma 1

Proof. Define the reward-tilted distribution

$$q(x) = \frac{1}{Z_T} p_{\mathrm{ref}}(x) \exp\left(\frac{r(x)}{\beta}\right), \quad Z_T = \int p_{\mathrm{ref}}(x) \exp\left(\frac{r(x)}{\beta}\right) dx. \tag{36}$$

Because r is bounded and p_{ref} is a probability distribution, $0 < Z_T < \infty$, and hence q is a well-defined probability distribution. Moreover, q and p_{ref} have the same support.

For any probability distribution p satisfying $p \ll p_{\mathrm{ref}}$, the KL-regularized reward objective can be rewritten as

$$\begin{aligned}\mathbb{E}_{x\sim p}[r(x)] - \beta\,\mathrm{KL}\left(p \,\|\, p_{\mathrm{ref}}\right) &= \int p(x) \left[r(x) - \beta \log \frac{p(x)}{p_{\mathrm{ref}}(x)} \right] dx \\ &= -\beta \int p(x) \log \frac{p(x)}{p_{\mathrm{ref}}(x) \exp\left(r(x)/\beta\right)} dx \\ &= -\beta \int p(x) \log \frac{p(x)}{Z_T q(x)} dx \\ &= -\beta \int p(x) \log \frac{p(x)}{q(x)} dx + \beta \log Z_T \\ &= \beta \log Z_T - \beta\,\mathrm{KL}\left(p \,\|\, q\right).\end{aligned} \tag{37}$$

Since $\mathrm{KL}(p\|q) \geq 0$, with equality if and only if $p = q$ almost everywhere, the objective is uniquely maximized by $p_T = q$. Therefore,

$$p_T(x) = \frac{1}{Z_T} p_{\mathrm{ref}}(x) \exp\left(\frac{r(x)}{\beta}\right), \tag{38}$$

which proves the result. □

B.4 Proof of Proposition 2

B.4.1 Optimal Extrapolated Distribution

Proof. For compactness, define the teacher-reference log-density ratio $g(x) = \log \frac{p_T(x)}{p_{\mathrm{ref}}(x)}$ on the support of p_{ref}. The distributional objective can then be written as

$$\begin{aligned}\mathcal{J}_\lambda(p) &= \lambda \int p(x) g(x)\, dx - \int p(x) \log \frac{p(x)}{p_{\mathrm{ref}}(x)} dx \\ &= -\int p(x) \log \frac{p(x)}{p_{\mathrm{ref}}(x) \exp\left(\lambda g(x)\right)} dx.\end{aligned} \tag{39}$$

Define

$$\begin{aligned}Z_\lambda &= \int p_{\mathrm{ref}}(x) \exp\left(\lambda g(x)\right) dx = \int p_{\mathrm{ref}}(x) \left(\frac{p_T(x)}{p_{\mathrm{ref}}(x)}\right)^\lambda dx, \\ p_\lambda(x) &= \frac{1}{Z_\lambda} p_{\mathrm{ref}}(x) \exp\left(\lambda g(x)\right) = \frac{1}{Z_\lambda} p_T(x)^\lambda p_{\mathrm{ref}}(x)^{1-\lambda}.\end{aligned} \tag{40}$$

By assumption, Z_λ is finite, so p_λ is well defined. Substituting $p_{\mathrm{ref}}(x)\exp(\lambda g(x)) = Z_\lambda p_\lambda(x)$ into Eq. (39) yields

$$\mathcal{J}_\lambda(p) = -\int p(x)\log\frac{p(x)}{Z_\lambda p_\lambda(x)}dx = -\int p(x)\log\frac{p(x)}{p_\lambda(x)}dx + \log Z_\lambda = \log Z_\lambda - \mathrm{KL}\left(p \,\|\, p_\lambda\right). \tag{41}$$

Since the KL divergence is nonnegative and vanishes if and only if $p = p_\lambda$ almost everywhere, p_λ is the unique optimizer of $\mathcal{J}_\lambda(p)$. □

B.4.2 Monotonicity of the Expected Reward

Proof. Under the reward-tilt assumption in Lemma 1, the teacher-reference log-density ratio satisfies

$$g(x) = \log\frac{p_T(x)}{p_{\mathrm{ref}}(x)} = \frac{r(x)}{\beta} - \log Z_T. \tag{42}$$

Substituting Eq. (42) into Eq. (40) gives

$$\begin{aligned} p_\lambda(x) &= \frac{p_{\mathrm{ref}}(x)\exp\left(\lambda g(x)\right)}{\int p_{\mathrm{ref}}(x')\exp\left(\lambda g(x')\right)dx'} \\ &= \frac{p_{\mathrm{ref}}(x)\exp\left(\lambda r(x)/\beta\right)\exp\left(-\lambda\log Z_T\right)}{\int p_{\mathrm{ref}}(x')\exp\left(\lambda r(x')/\beta\right)\exp\left(-\lambda\log Z_T\right)dx'} \\ &= \frac{p_{\mathrm{ref}}(x)\exp\left(\lambda r(x)/\beta\right)}{\widetilde{Z}_\lambda}, \\ \widetilde{Z}_\lambda &= \int p_{\mathrm{ref}}(x)\exp\left(\lambda r(x)/\beta\right)dx. \end{aligned} \tag{43}$$

Because r is bounded, differentiation under the integral sign is valid. Differentiating the log-density of p_λ with respect to λ yields

$$\begin{aligned} \frac{\partial}{\partial\lambda}\log p_\lambda(x) &= \frac{r(x)}{\beta} - \frac{\partial}{\partial\lambda}\log\widetilde{Z}_\lambda \\ &= \frac{r(x)}{\beta} - \frac{1}{\widetilde{Z}_\lambda}\int p_{\mathrm{ref}}(x')\exp\left(\frac{\lambda r(x')}{\beta}\right)\frac{r(x')}{\beta}dx' \\ &= \frac{1}{\beta}\left(r(x) - \mathbb{E}_{x'\sim p_\lambda}[r(x')]\right) \\ &= \frac{1}{\beta}\left(r(x) - J(\lambda)\right). \end{aligned} \tag{44}$$

Therefore,

$$\begin{aligned} \frac{dJ(\lambda)}{d\lambda} &= \frac{d}{d\lambda}\int p_\lambda(x)r(x)\,dx \\ &= \int r(x)\frac{\partial p_\lambda(x)}{\partial\lambda}dx \\ &= \int r(x)p_\lambda(x)\frac{\partial\log p_\lambda(x)}{\partial\lambda}dx \\ &= \frac{1}{\beta}\int p_\lambda(x)r(x)\left(r(x) - J(\lambda)\right)dx \\ &= \frac{1}{\beta}\left(\mathbb{E}_{x\sim p_\lambda}[r(x)^2] - \mathbb{E}_{x\sim p_\lambda}[r(x)]^2\right) \\ &= \frac{1}{\beta}\mathrm{Var}_{x\sim p_\lambda}[r(x)] \geq 0. \end{aligned} \tag{45}$$

The inequality is strict exactly when $\mathrm{Var}_{p_\lambda}[r(x)] > 0$, or equivalently, when r is not constant p_λ-almost surely. This proves the claimed monotonicity. □

B.5 Proof of Proposition 3

Recall the reward-aligned distribution family

$$p_\rho(x) = \frac{1}{Z_\rho} p_0(x) \exp\left(\frac{\rho r(x)}{\beta}\right), \quad Z_\rho = \int p_0(x) \exp\left(\frac{\rho r(x)}{\beta}\right) dx. \tag{46}$$

In particular, the teacher is given by

$$p_T(x) = p_1(x) = \frac{1}{Z_1} p_0(x) \exp\left(\frac{r(x)}{\beta}\right). \tag{47}$$

Because r is bounded, all normalizing constants appearing below are finite and strictly positive.

B.5.1 Effective Reward-Tilt Coefficient

Proof. Using $p_T = p_1$ and $p_{\text{ref}} = p_\rho$, the unnormalized extrapolated distribution can be expanded as

$$\begin{aligned} p_T(x)^\lambda p_\rho(x)^{1-\lambda} &= \left[\frac{1}{Z_1} p_0(x) \exp\left(\frac{r(x)}{\beta}\right)\right]^\lambda \left[\frac{1}{Z_\rho} p_0(x) \exp\left(\frac{\rho r(x)}{\beta}\right)\right]^{1-\lambda} \\ &= Z_1^{-\lambda} Z_\rho^{\lambda-1} p_0(x)^{\lambda+1-\lambda} \exp\left(\frac{\lambda r(x) + (1-\lambda)\rho r(x)}{\beta}\right) \\ &= Z_1^{-\lambda} Z_\rho^{\lambda-1} p_0(x) \exp\left(\frac{[\lambda + (1-\lambda)\rho] r(x)}{\beta}\right) \\ &= Z_1^{-\lambda} Z_\rho^{\lambda-1} p_0(x) \exp\left(\frac{[1 + (\lambda-1)(1-\rho)] r(x)}{\beta}\right). \end{aligned} \tag{48}$$

The factor $Z_1^{-\lambda} Z_\rho^{\lambda-1}$ is independent of x and is therefore absorbed into the normalizing constant. Defining

$$\rho_{\text{eff}}(\lambda, \rho) = 1 + (\lambda - 1)(1 - \rho), \quad Z_{\lambda,\rho} = \int p_0(x) \exp\left(\frac{\rho_{\text{eff}}(\lambda, \rho) r(x)}{\beta}\right) dx, \tag{49}$$

the normalized extrapolated distribution is

$$p_{\lambda,\rho}(x) = \frac{1}{Z_{\lambda,\rho}} p_0(x) \exp\left(\frac{\rho_{\text{eff}}(\lambda, \rho) r(x)}{\beta}\right). \tag{50}$$

Thus, $p_{\lambda,\rho}$ remains in the same reward-aligned exponential family, with effective alignment coefficient $\rho_{\text{eff}}(\lambda, \rho)$.

Now consider two references satisfying $\rho_d < \rho_n < 1$ and fix $\lambda > 1$. Since $\lambda - 1 > 0$, we obtain

$$\begin{aligned} \rho_{\text{eff}}(\lambda, \rho_d) - \rho_{\text{eff}}(\lambda, \rho_n) = (\lambda - 1)\left[(1 - \rho_d) - (1 - \rho_n)\right] = (\lambda - 1)(\rho_n - \rho_d) > 0, \\ \rho_{\text{eff}}(\lambda, \rho_n) - 1 = (\lambda - 1)(1 - \rho_n) > 0. \end{aligned} \tag{51}$$

Consequently,

$$\rho_{\text{eff}}(\lambda, \rho_d) > \rho_{\text{eff}}(\lambda, \rho_n) > 1, \tag{52}$$

which proves the first part of the proposition. □

B.5.2 Local Reward Sensitivity

Proof. For a fixed reference parameter ρ, define

$$\begin{aligned} J_\rho(\lambda) &= \mathbb{E}_{x \sim p_{\lambda,\rho}}[r(x)], \\ p_{\lambda,\rho}(x) &= \frac{1}{Z_{\lambda,\rho}} p_0(x) \exp\left(\frac{\rho_{\text{eff}}(\lambda, \rho) r(x)}{\beta}\right), \quad \rho_{\text{eff}}(\lambda, \rho) = 1 + (\lambda - 1)(1 - \rho). \end{aligned} \tag{53}$$

The effective alignment coefficient satisfies

$$\frac{\partial \rho_{\text{eff}}(\lambda, \rho)}{\partial \lambda} = 1 - \rho. \tag{54}$$

Because r is bounded, differentiation under the integral sign is valid. Differentiating the log normalizer gives

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log Z_{\lambda,\rho} &= \frac{1}{Z_{\lambda,\rho}}\int p_0(x)\exp\left(\frac{\rho_{\text{eff}}(\lambda,\rho)r(x)}{\beta}\right)\frac{r(x)}{\beta}\frac{\partial\rho_{\text{eff}}(\lambda,\rho)}{\partial\lambda}dx \\ &= \frac{1-\rho}{\beta}\int p_{\lambda,\rho}(x)r(x)\,dx \\ &= \frac{1-\rho}{\beta}J_\rho(\lambda).\end{aligned} \tag{55}$$

It follows that the derivative of the log density is

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log p_{\lambda,\rho}(x) &= \frac{r(x)}{\beta}\frac{\partial\rho_{\text{eff}}(\lambda,\rho)}{\partial\lambda} - \frac{\partial}{\partial\lambda}\log Z_{\lambda,\rho} \\ &= \frac{1-\rho}{\beta}r(x) - \frac{1-\rho}{\beta}J_\rho(\lambda) \\ &= \frac{1-\rho}{\beta}\left[r(x) - J_\rho(\lambda)\right].\end{aligned} \tag{56}$$

The reward sensitivity with respect to λ is therefore

$$\begin{aligned}\frac{dJ_\rho(\lambda)}{d\lambda} &= \frac{d}{d\lambda}\int p_{\lambda,\rho}(x)r(x)\,dx \\ &= \int r(x)\frac{\partial p_{\lambda,\rho}(x)}{\partial\lambda}dx \\ &= \int p_{\lambda,\rho}(x)r(x)\frac{\partial}{\partial\lambda}\log p_{\lambda,\rho}(x)dx \\ &= \frac{1-\rho}{\beta}\int p_{\lambda,\rho}(x)r(x)\left[r(x) - J_\rho(\lambda)\right]dx \\ &= \frac{1-\rho}{\beta}\left(\mathbb{E}_{p_{\lambda,\rho}}[r(x)^2] - \mathbb{E}_{p_{\lambda,\rho}}[r(x)]^2\right) \\ &= \frac{1-\rho}{\beta}\text{Var}_{x\sim p_{\lambda,\rho}}[r(x)].\end{aligned} \tag{57}$$

At $\lambda = 1$, $p_{1,\rho}(x) = \frac{1}{Z_1}p_0(x)\exp\left(\frac{r(x)}{\beta}\right) = p_T(x)$. The effective alignment coefficient is independent of the reference. Hence, all reference choices recover the same teacher distribution at $\lambda = 1$, and Eq. (57) reduces to

$$\left.\frac{dJ_\rho(\lambda)}{d\lambda}\right|_{\lambda=1} = \frac{1-\rho}{\beta}\text{Var}_{x\sim p_T}[r(x)]. \tag{58}$$

Because r is nonconstant under p_T, $\text{Var}_{p_T}[r(x)] > 0$. For $\rho_d < \rho_n < 1$, both local derivatives are therefore strictly positive, and their ratio is

$$\left.\frac{dJ_{\rho_d}(\lambda)/d\lambda}{dJ_{\rho_n}(\lambda)/d\lambda}\right|_{\lambda=1} = \frac{\frac{1-\rho_d}{\beta}\text{Var}_{p_T}[r(x)]}{\frac{1-\rho_n}{\beta}\text{Var}_{p_T}[r(x)]} = \frac{1-\rho_d}{1-\rho_n} > 1, \tag{59}$$

where the final inequality follows from $\rho_d < \rho_n < 1$. This proves the local reward-sensitivity claim. □

C Training Algorithm

We provide the detailed training algorithm in Algorithm 1.

D Experimental Details

D.1 Training Configuration

All experiments are conducted on a single node with eight NVIDIA A100 GPUs. We follow the experimental configurations of DiffusionOPD (Li et al., 2026b). All experiments use **`SD3.5-M`** at a

Algorithm 1 `DreOPD`

Require: teachers $\{v_T^{(m)}\}_{m=1}^M$; degraded reference v_{deg}; student v_θ; extrapolation strength $\lambda > 1$; tasks $\mathcal{M} = \{1, \dots, M\}$; prompt datasets $\{\mathcal{C}^{(m)}\}_{m=1}^M$

for each training round **do**
  for $m = 1, \dots, M$ **do**
    Sample prompts $c \sim \mathcal{C}^{(m)}$
    Rollout the student v_θ on c to obtain trajectory $\{x_{t_j}\}_{j=0}^N$
    Query $v_T^{(m)}(x_{t_j}, t_j, c)$ and $v_{\text{deg}}(x_{t_j}, t_j, c)$ for $j \in \{0, \dots, N-1\}$ ▷ no grad
    Compute $\mathcal{L}_m(\theta)$ in Eq. (13)
    Update θ by one optimizer step on $\mathcal{L}_m(\theta)$
  end for
end for

resolution of 512×512. We fine-tune LoRA adapters with rank $r = 32$ and scaling factor $\alpha = 64$, while keeping the teacher and reference models frozen. We optimize the LoRA parameters using AdamW with a learning rate of 3×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}, and $\epsilon = 10^{-8}$. Student trajectories are collected using a 10-step first-order ODE sampler, and evaluation uses 40 sampling steps. Unless otherwise specified, the remaining optimization, sampling, and task-specific configurations follow the corresponding baseline settings.

D.2 Teachers and Baselines

All methods are initialized from the same **`SD3.5-M`** (Esser et al., 2024) checkpoint. For OPD-based methods, the task-specific teachers remain frozen throughout training and are queried only at states visited by the current student. Unless otherwise specified, we follow the teacher construction and baseline configurations adopted by DiffusionOPD (Li et al., 2026b). We summarize the relevant details below.

Single-task RL teachers. We use three independently trained teachers, each specialized for one of the target capabilities: compositional prompt following, text rendering, and aesthetic quality. These are the same task-specific teachers used by DiffusionOPD (Li et al., 2026b).

The **`GenEval teacher`** is trained using DiffusionNFT (Zheng et al., 2025b) to optimize the GenEval compositional reward. GenEval evaluates compositional prompt following through rule-based correctness signals covering object identity, counting, color, spatial relations, and attribute binding. The **`OCR teacher`** is trained using GRPO-Guard (Wang et al., 2026) to optimize the text-rendering reward. The **`Aesthetics teacher`** is also trained using GRPO-Guard and optimizes the equally weighted reward of PickScore, ClipScore and HPSv2.1.

Each teacher is trained only on its corresponding prompt distribution and reward objective. Consequently, the reported **`Teacher`** results in the single-teacher experiment combine the in-domain scores of three different specialized models and do not represent a single multi-task teacher.

Multi-task RL baselines. We compare against direct multi-task optimization with Flow-GRPO, GRPO-Guard, DiffusionNFT, and CascadeNFT. These baselines receive the same task-specific prompt datasets and reward definitions used to construct the teachers, but optimize a single shared model rather than training separate specialists.

For **`Flow-GRPO`** (Liu et al., 2026), the reverse-time flow sampler is treated as a stochastic policy with Gaussian transition kernels, and the model is optimized using group-relative policy updates. In the multi-task setting, training alternates among the GenEval, OCR, and aesthetics prompt datasets, applying the reward associated with the sampled task. **`GRPO-Guard`** (Wang et al., 2026) follows the same alternating multi-task organization while incorporating its guarded optimization mechanism to improve training stability under potentially exploitable reward signals.

`DiffusionNFT` (Zheng et al., 2025b) performs direct reward-based fine-tuning through the differentiable generation process. Its multi-task variant alternates among the three task datasets and

applies the corresponding reward objective at each update. We use the same task curriculum across jointly trained multi-task baselines.

The original multi-task formulation of DiffusionNFT adopts sequential optimization. We denote this variant as **`CascadeNFT`**. The model is first fine-tuned on one task and then successively adapted to the remaining tasks using their respective objectives.

OPD baselines. To enable a fair comparison, all OPD methods use the same student initialization, task-specific teachers, prompt datasets, and task-sampling schedule.

`Flow-OPD` (Fang et al., 2026) regresses the student velocity toward the selected teacher velocity on student-generated states. It additionally employs Manifold Anchor Regularization to constrain the student toward a high-quality visual manifold.

`DiffusionOPD` (Li et al., 2026b) minimizes the stepwise reverse KL divergence from the student transition to the corresponding teacher transition along student-generated trajectories.

D.3 Rewards and Metrics

We evaluate the models using the following metrics:

- **GenEval** (Ghosh et al., 2023) measures compositional prompt following, including object identity, counting, color, spatial relations, and attribute binding.
- **OCR** measures the accuracy of text rendered in generated images using the prompt split released with Flow-GRPO (Liu et al., 2026). Here, we perform text recognition using `PaddleOCR=3.7.0` with `PaddlePaddle-GPU=3.3.1`.
- **PickScore** (Kirstain et al., 2023) is a preference-based metric trained to assess human preferences for text-to-image generations.
- **ClipScore** (Hessel et al., 2021) measures the semantic correspondence between a generated image and its text prompt using CLIP representations.
- **HPSv2.1** (Wu et al., 2023) evaluates text-image alignment and perceptual quality using a model trained on human preference data.
- **Aesthetic Score** (Schuhmann, 2022) estimates the perceptual quality of generated images. We use it as an out-of-domain metric because it is not directly included in the training reward.
- **ImageReward** (Xu et al., 2023) evaluates human preference and text-image alignment. It is also used only for out-of-domain evaluation.

Average score. Because the metrics have different numerical ranges, we compute the average score by independently applying min-max normalization to each metric. Let $s_{m,k}$ denote the score of method m on metric k, and let $\mathcal{M}$ denote the set of methods included in the corresponding table. We compute

$$\widetilde{s}_{m,k} = \frac{s_{m,k} - s_k^{\min}}{s_k^{\max} - s_k^{\min}}, \quad \mathrm{Avg}(m) = \frac{1}{K} \sum_{k=1}^{K} \widetilde{s}_{m,k}, \tag{60}$$

where K is the number of metrics included in the table. All metrics are positively oriented, such that a larger value indicates better performance.

D.4 Degraded-Reference Construction

The reference is the pretrained **`SD3.5-M`** backbone without LoRA adapters and with classifier-free guidance disabled. We keep the sampler, noise schedule, and guidance scale fixed across all degradation variants and modify only the reference velocity or backbone weights.

Velocity Quantization. We apply per-channel quantization directly to the reference velocity output using either 4-bit or 8-bit precision. This operation introduces deterministic quantization error without modifying the backbone parameters. We use 8-bit velocity quantization in the main experiments, as it achieves the best average performance in our ablation.

Weight Quantization. We apply per-output-channel quantization to all multi-dimensional weight tensors in the frozen reference backbone, while excluding normalization parameters and biases. We evaluate both 4-bit and 8-bit variants. Unlike velocity quantization, this degradation affects intermediate representations throughout the denoising network.

Gaussian Velocity Perturbation. We add zero-mean Gaussian noise to the reference velocity output, with the noise scale set relative to the empirical standard deviation of each channel. We evaluate relative noise levels of 0.01 and 0.1. This construction provides a stochastic alternative to the deterministic perturbations introduced by quantization.

Reference	GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImageReward
Original reference	0.2529	0.1377	20.519	0.2384	0.2052	5.1614	-0.5471
4-bit velocity quantization	0.1939↓	0.1292↓	20.540↑	0.2345↓	0.2118↑	5.1908↑	-0.5256↑
8-bit velocity quantization	0.2501↓	0.1327↓	20.550↑	0.2383↓	0.2064↑	5.1162↓	-0.5457↑
4-bit weight quantization	0.0081↓	0.0112↓	18.312↓	0.1644↓	0.0964↓	4.0362↓	-2.2245↓
8-bit weight quantization	0.2371↓	0.1295↓	20.519=	0.2383↓	0.2049↓	5.1252↓	-0.5881↓
Gaussian perturbation ($\sigma = 0.01$)	0.2309↓	0.1357↓	20.550↑	0.2397↑	0.2059↑	5.1288↓	-0.5689↓
Gaussian perturbation ($\sigma = 0.1$)	0.2514↓	0.1255↓	20.574↑	0.2394↑	0.2072↑	5.1189↓	-0.5343↑

Table 5: Performance of the original and degraded references across all evaluated reward metrics.

Performance of the Degraded References Table 5 reports the reward metrics of the original and degraded references. All degradation variants reduce GenEval and OCR performance, indicating weaker alignment with these two rule-based objectives. The changes in the preference-based metrics are less consistent. Several variants slightly improve PickScore, ClipScore, HPSv2.1, or ImageReward while reducing other scores. Reference degradation therefore does not uniformly lower every reward. Instead, it changes the reward profile and increases the contrast with the task-specific teachers.

The extent of these changes depends on the degradation mechanism. 4-bit weight quantization substantially reduces every metric and may disrupt the generative structure of the reference. 8-bit velocity quantization produces a milder change. It reduces GenEval, OCR ClipScore and Aesthetic while preserving the remaining metrics. This balance provides a meaningful teacher-reference contrast without the broad performance degradation caused by aggressive weight quantization and motivates its use in our main experiments.

Gaussian perturbations inject unstructured random noise into the velocity field and may disrupt the underlying generation structure. Quantization instead introduces bounded and deterministic discretization errors that better preserve the structure of the reference predictions.

This observation also clarifies the scope of Proposition 3, which assumes that the teacher and reference belong to the same reward-tilted family p_ρ. The scores in Table 5 neither establish nor rule out this assumption, since plausible generation behavior or changes in individual metrics do not imply the required exponential-tilt relationship. We therefore interpret the proposition as an idealized account of contrast amplification under structured reward alignment, while evaluating practical reference perturbations empirically. The broad performance degradation caused by 4-bit weight quantization suggests that aggressive perturbations may introduce corruption directions unrelated to the task reward, making the resulting teacher-reference contrast less reliable for extrapolation.

E Additional Experimental Results

E.1 Complete Single-Teacher Results

Table 6 reports the complete results for the three independently distilled students. Each student is optimized using one task-specific teacher.

For **`GenEval Teacher`** distillation, Our **`DreOPD`** improves the teacher score from 0.9470 to 0.9710 and achieves the best GenEval result among the compared methods. This stronger specialization does not improve every secondary metric. In particular, OCR and several perceptual scores decrease relative to the GenEval teacher. The result indicates that extrapolation strengthens the target capability while retaining the trade-offs associated with single-task optimization.

	Model	In-Domain Reward					OOD Reward	
		GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImgRwd
Base	SD3.5-M (w/o CFG)	0.2529	0.1377	20.519	0.2384	0.2052	5.161	-0.5471
	SD3.5-M	0.6273	0.5079	22.331	0.2837	0.2795	5.396	0.8324
GenEval	GenEval Teacher	0.9470	0.6286	20.084	0.2870	0.2644	5.246	0.8976
	Flow-OPD	0.9501$_{+0.33\%}$	0.6077	22.161	0.2881	0.2673	5.270	0.9507
	DiffusionOPD	0.9648$_{+1.88\%}$	0.4641	22.126	0.2794	0.2447	5.169	0.6209
	Ours	**0.9710**$_{+2.53\%}$	0.4748	22.209	0.2766	0.2446	5.179	0.5873
OCR	OCR Teacher	0.6562	0.9239	22.225	0.2919	0.2720	5.266	0.8881
	Flow-OPD	0.6556	0.9279$_{+0.43\%}$	22.227	0.2905	0.2719	5.267	0.9029
	DiffusionOPD	0.6533	0.9246$_{+0.08\%}$	22.199	0.2905	0.2709	5.264	0.8721
	Ours	0.6638	**0.9364**$_{+1.53\%}$	22.158	0.2911	0.2719	5.279	0.9007
Aes.	Aes Teacher	0.4935	0.5014	<u>24.034</u>	**0.2963**	<u>0.3460</u>	6.232	1.5071
	Flow-OPD	0.4758	0.4913	24.006$_{-0.12\%}$	<u>0.2957</u>$_{-0.20\%}$	0.3445$_{-0.43\%}$	6.216	1.5027
	DiffusionOPD	0.4738	0.4914	24.008$_{-0.11\%}$	0.2955$_{-0.27\%}$	0.3454$_{-0.17\%}$	6.233	1.5030
	Ours	0.4739	0.4730	**24.037**$_{+0.01\%}$	0.2955$_{-0.27\%}$	**0.3497**$_{+1.07\%}$	**6.301**	**1.5199**

Table 6: Complete results of single-task OPD methods. **Bold**: best; <u>Underline</u>: second best; Gray-colored: In-Domain reward. Subscripts show relative changes from the corresponding teacher.

For **OCR Teacher** distillation, Our **DreOPD** improves OCR from 0.9239 to 0.9364. It also slightly improves GenEval, Aesthetic Score, and ImageReward over the OCR teacher while keeping the remaining preference metrics close to their teacher values.

For **Aes Teacher** distillation, Our **DreOPD** improves PickScore from 24.034 to 24.037 and HPSv2.1 from 0.3460 to 0.3497, while ClipScore decreases slightly from 0.2963 to 0.2955. The same student also improves the out-of-domain Aesthetic Score from 6.232 to 6.301 and ImageReward from 1.5071 to 1.5199.

Across all three settings, **DreOPD** achieves the largest improvement on the task-specific metrics. Some off-task metrics decrease as the student becomes more specialized, although these changes are generally moderate when considered across the full set of metrics. The overall results therefore indicate a favorable balance between improving the target capability and retaining the remaining capabilities.

E.2 Comparison with Training-Free Model Merging

We further compare **DreOPD** with training-free model merging baselines. Given three task-specific teachers, we evaluate standard weight averaging, which directly merges their parameters without additional training. We also consider ExPO weight extrapolation (Zheng et al., 2025a), where the averaged teacher model is extrapolated with $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ following the model extrapolation formulation. These training-free methods are simple and efficient, but they combine teachers only in parameter space and do not optimize the merged model on generation trajectories. As a result, weight averaging may dilute specialized capabilities, while weight extrapolation may move along directions that are not aligned with generation quality. In contrast, **DreOPD** performs on-policy distillation with extrapolative velocity targets. As shown in Table 7, this leads to stronger and more balanced performance across evaluation metrics than training-free merging baselines.

F Prompts for Text-to-Image Generation

Below we list the prompts for text-to-image generation in Figure 1.

- A gray tabby cat sitting on a matte concrete floor against a plain pale gray studio wall.
- A single ripe strawberry hanging from its plant, soft cloudy daylight, plain softly blurred dark green leaf background, individual seeds and glossy red skin clearly resolved, tiny water droplets on the surface, 100mm macro, realistic botanical photograph.
- A single wooden pier extending into a calm lake at dawn, glassy water, mist over the surface, realistic landscape photograph.

Model	GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImgReward
`SD3.5-M`	0.6273	0.5079	22.331	0.2837	0.2795	5.396	0.8324
Teachers							
`GenEval Teacher`	0.9470	0.6286	20.084	0.2870	0.2644	5.246	0.8976
`OCR Teacher`	0.6562	0.9239	22.225	0.2919	0.2720	5.266	0.8881
`Aes Teacher`	0.4935	0.5014	24.034	0.2963	0.3460	6.232	1.5071
Model Merge							
`Weight Average`	0.8019	0.6952	22.761	0.2915	0.2965	5.439	1.1509
`ExPO` $\alpha = 0.25$	0.8235	0.7304	22.805	0.2923	0.2983	5.440	1.2034
`ExPO` $\alpha = 0.5$	0.8465	0.7670	22.828	0.2941	0.2999	5.449	1.2379
`ExPO` $\alpha = 0.75$	0.8569	0.7901	22.850	0.2951	0.3006	5.439	1.2654
`ExPO` $\alpha = 1.0$	0.8685	0.7994	22.856	0.2958	0.3014	5.443	1.2859
`Ours (w/o DeRef)`	0.9668	0.9281	24.032	0.2956	**0.3487**	6.239	1.5172
`Ours`	**0.9681**	0.9362	**24.035**	0.2959	**0.3487**	**6.292**	**1.5245**

Table 7: Comparison with training-free model merging methods using three task-specific teachers. **Bold**: best; Underline: second best; `DeRef`: degraded reference; Gray-colored: In-Domain reward; Blue-colored: Ours.

- A reading-nook flat-lay on a plain matte pale-gray tabletop: an open hardcover book in the center, a folded charcoal-gray wool blanket in the upper left, a matte-white porcelain cup of tea in the upper right, and a small pair of round reading glasses in the lower center, soft cool north-window light, uncluttered pale background, top-down 50mm f/4, realistic photograph.
- A bowl of noodles on a wooden table.
- An anime girl on the seaside bathed in morning light.
- A deer standing in a foggy forest.
- A tea ceremony still-life on a plain matte pale-gray slate tray: a small cast-iron teapot in the center, a small ceramic bowl of loose green tea leaves in the upper left, three tiny porcelain tea cups arranged in an arc below the teapot, and a bamboo whisk in the upper right, soft cool diffused daylight, uncluttered pale background, top-down 50mm f/4, realistic photograph.
- A great horned owl perched on a bare weathered branch against a smooth deep-blue twilight sky.
- A weathered enamel sign reading "FRESH BREAD DAILY" hanging above a plain oak bakery counter, uncluttered warm-gray plaster wall background, tiny paint chips on the enamel.
- A cinematic close-up portrait of an elderly fisherman with a weathered face and kind eyes, soft diffused window light, shallow depth of field.
- A watercolor painting of cherry blossom trees beside a quiet river.

G Additional Qualitative Results

Figure 6 to Figure 8 presents additional examples over prompt-following, text-rendering and aesthetic generation.

H Future Directions

Future work may explore more adaptive ways to construct degraded references and schedule the extrapolation strength during training. It would also be valuable to extend `DreOPD` beyond text-to-image generation, such as video generation and controllable image editing, to further examine the generality of degraded-reference extrapolative distillation.



Figure 6: Additional qualitative results on aesthetic generation. The examples cover natural scenes, artistic styles, and complex compositions.

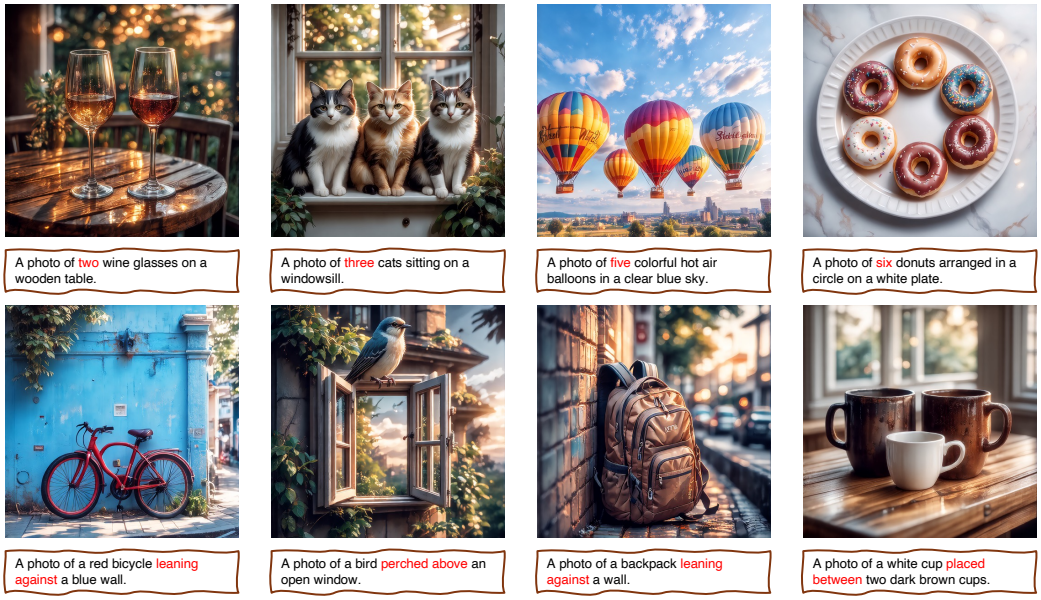


Figure 7: Additional qualitative results on compositional prompt following. The examples cover counting, color recognition, spatial relations, and attribute binding.



Figure 8: Additional qualitative results on text rendering. The examples include short text, long text an complex backgrounds.