D-OPD: ENHANCING INTER-TEACHER DISCRIMINABILITY IN MULTI-TASK ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

For text-to-image (T2I) generation, existing reinforcement learning (RL) methods are constrained by reward hacking and gradient conflicts across different objectives, leading to poor performance in multi-task scenarios. Although the recent On-Policy Distillation (OPD) paradigm mitigates these challenges by constructing dense supervision signals, significant room for improvement remains. In particular, vanilla OPD merely enforces the alignment between the student model and the task-specific teacher without constraining its distributional divergence from the remaining ones. This prevents the model from disentangling the knowledge of irrelevant teachers when optimizing for a given task. To address this issue, we propose **D-OPD**, an efficient framework that reformulates the multi-teacher on-policy distillation as a task-expert classification problem. Concretely, we first derive a categorical distribution over teachers based on the student–teacher discrepancy, quantifying the likelihood of the student's rollout states being attributed to each teacher. By improving the probability of the task-specific teacher via a cross-entropy objective, D-OPD pulls the student model closer to the corresponding task expert, while maintaining distributional separation from the rest. Benefiting from this, D-OPD significantly enhances the *inter-teacher discriminability* of the student's rollouts, facilitating faster convergence and superior performance. In addition, D-OPD incorporates *Dual-Teacher* (DT) and *Early-Window* (EW) strategies to alleviate the computational overhead incurred by an increasing number of teachers, enhancing training efficiency without compromising the overall performance. Extensive experiments demonstrate that D-OPD consistently outperforms RL and OPD baselines across various settings, highlighting its significant promise for multi-task post-training.

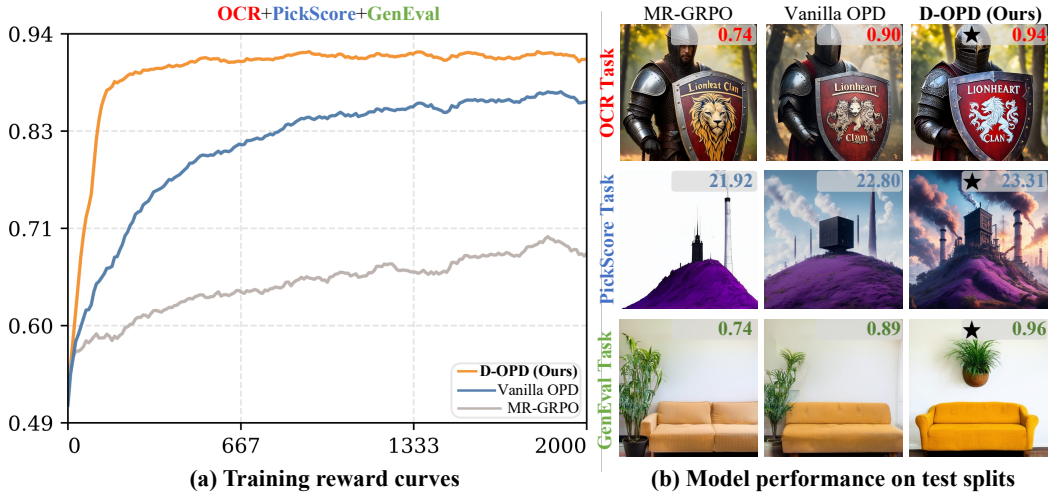
1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a dominant post-training paradigm in both large language models (LLMs) Chen et al. (2024); Wang et al. (2024) and text-to-image (T2I) generation Podell et al. (2023); Esser et al. (2024); Rombach et al. (2022); Labs (2024); Ho et al. (2020), aligning the AI-generated content with human preferences. However, existing RL methods Liu et al. (2025); Guo et al. (2025); Wallace et al. (2024); Rafailov et al. (2023) merely leverage the sparse scalar reward derived from the final model output, making them susceptible to reward hacking. Moreover, the lack of dense guidance exacerbates gradient conflicts among different objectives, leading to suboptimal performance in multi-task scenarios. Consequently, designing a more efficient post-training algorithm remains a significant challenge.

Inspired by recent efforts in LLMs Li et al. (2026b); Zhang et al. (2026); Lu & Thinking Machines Lab (2025), Flow-OPD Fang et al. (2026) and Diffusion-OPD Li et al. (2026a) have introduced On-Policy Distillation (OPD) into the T2I generation domain, alleviating the above challenges. In contrast to the RL counterparts Liu et al. (2025); Zheng et al. (2025), OPD provides dense supervision signals derived from teacher models for step-wise guidance, effectively mitigating reward hacking. As indicated by previous literature Li et al. (2026a); Fang et al. (2026), OPD significantly outperforms multi-task RL baselines in overall performance.

Despite the encouraging advances, OPD in T2I generation is still at an early stage, leaving substantial room for improvement. In particular, existing implementations merely enforce the alignment

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.



(a) Training reward curves **(b) Model performance on test splits**

Figure 1: The figures compare our D-OPD with Multi-Reward GRPO (MR-GRPO) and vanilla OPD in a multi-task setting, including OCR Du et al. (2020), PickScore Kirstain et al. (2023), and GenEval Ghosh et al. (2023). The training reward curves illustrated in **(a)** indicate that our method achieves significantly faster convergence. Besides, **(b)** demonstrates that D-OPD outperforms other methods in performance for each independent task.

between the student and the task expert, without constraining the student's distributional discrepancies from other irrelevant teachers. To illustrate this limitation, we first construct a multi-teacher categorical distribution based on the student-teacher discrepancy. Specifically, a higher probability indicates that the student's divergence from the corresponding teacher is significantly smaller than from others, implying that the student's rollout states are more likely to be attributed to this teacher for a given task. As demonstrated in Fig. 2 (b), the probability disparity between the task expert and irrelevant teachers remains limited under the vanilla OPD objective. This indicates that the student's own rollouts are indistinguishable across teachers, resulting in low *inter-teacher discriminability*. This issue is particularly pronounced in high-noise rollout steps, which determine the global content of the final outcome, such as coarse geometric and semantic structure. Therefore, the student model fails to effectively disentangle extraneous knowledge from irrelevant teachers when optimizing for a given task, resulting in suboptimal training efficiency and performance ceiling.

To alleviate the challenge, we propose **D-OPD**, a simple yet effective framework designed to enhance OPD in multi-task scenarios. Crucially, it reformulates the multi-teacher on-policy distillation as a task-expert classification problem, enhancing the established probability of the task-specific teacher through an InfoNCE-style cross-entropy (CE) objective Oord et al. (2018). Benefiting from this, D-OPD effectively pulls the student closer to the task-specific teacher, while maintaining distributional separation from the rest. As illustrated in Fig. 2 (b), our method achieves a significant improvement in inter-teacher discriminability, improving both the convergence speed and the final performance, as evidenced in Fig. 1. Despite its effectiveness, the computational overhead becomes prohibitive as the number of teachers increases. To address this, D-OPD further introduces the *Dual-Teacher* (DT) strategy, which constructs the categorical distribution using only two teachers rather than the entire set, significantly reducing the computational complexity. Moreover, we find that enhancing inter-teacher discriminability yields the most substantial gains during the high-noise rollout stages. Motivated by this, we propose the *Early-Window* (EW) strategy, which applies the CE loss only in early steps and the vanilla OPD objective for the rest of the rollout. This mechanism further reduces computation and accelerates convergence without compromising the overall performance. In summary, our contributions are as follows:

1. The proposed D-OPD reformulates the multi-teacher on-policy distillation as a task-expert classification problem, improving the training efficiency and performance ceiling.
2. To mitigate the computational overhead incurred by multiple teachers, D-OPD incorporates DT and EW strategies, accelerating convergence and maintaining the final performance.
3. As evident in our experiments, D-OPD consistently outperforms RL and OPD baselines across various settings, demonstrating significant potential in multi-task scenarios.

2 Related Works

2.1 Reinforcement Learning from Human Feedback

In large language models (LLMs) Chen et al. (2024); Bai et al. (2025); Wang et al. (2024), reinforcement learning Rafailov et al. (2023); Guo et al. (2025) has significantly enhanced the alignment between model outputs and human preferences. Several studies have introduced these techniques into text-to-image (T2I) diffusion models Dhariwal & Nichol (2021); Ho et al. (2020). Early approaches Fan et al. (2023); Xu et al. (2023) directly maximize the final reward and backpropagate gradients throughout the diffusion trajectory. DPO Wallace et al. (2024) aligns the model outputs with the preferred samples while distancing them from inferior ones. GRPO Guo et al. (2025); Liu et al. (2025) eliminates the reliance on an explicit value network adopted in PPO-style methods Schulman et al. (2017), achieving a favorable trade-off between efficiency and performance. Despite significant advances, existing methods are constrained by sparse scalar rewards, leading to reward hacking and exacerbating gradient conflicts in multi-task scenarios.

2.2 On-Policy Distillation

In LLMs, On-Policy Distillation (OPD) Zhao et al. (2026); Lu & Thinking Machines Lab (2025); Agarwal et al. (2024); Gu et al. (2024); Ko et al. (2025); Yang et al. (2026) has been introduced to provide dense guidance, mitigating reward hacking in current RL methods. Specifically, it constructs the token-level supervision signals by calculating the discrepancy between the student and teacher models on the student's own rollout. Recently, Flow-OPD Fang et al. (2026) and its concurrent work, Diffusion-OPD Li et al. (2026a), have introduced this paradigm Jiang et al. (2026) to the post-training stage in T2I generation. Their experimental results demonstrate that OPD significantly outperforms RL baselines in multi-task optimization settings. Nevertheless, existing OPD implementation in T2I generation is still at an early stage, leaving substantial room for improvement.

3 Method

3.1 Preliminary: On-Policy Distillation

We first revisit the formulation of On-Policy Distillation (OPD) in single-task optimization. For flow matching models, OPD requires the student model to sample T-length rollouts conditioned on the prompt c, where the state at the j-th step is denoted as x_{t_j}. Then, it minimizes the reverse KL-divergence between the student and the teacher model at each step:

$$\mathcal{L}_{\text{OPD}}^{\text{KL}} = \mathbb{E}_{c\sim D, x_{t_{T:1}}\sim p^S_{\theta_{\text{old}}}(c)} \left[\sum_{j=T}^{1} \text{KL}\Big(p^S_\theta(\cdot \mid x_{t_j}, c) \,\Big\|\, p^T_\phi(\cdot \mid x_{t_j}, c) \Big) \right], \tag{1}$$

where D is the training dataset. $p^S_\theta(\cdot \mid x_{t_j}, c)$ and $p^T_\phi(\cdot \mid x_{t_j}, c)$ denote the state transition kernels of the student and teacher models at the j-th step, respectively. $p^S_{\theta_{\text{old}}}$ is the current policy used for sampling the trajectories. To derive the closed-form expression for the KL-divergence in Eq. (1), we first present the reverse Stochastic Differential Equation (SDE) for the flow matching model:

$$x_{t_j+\Delta t} = x_{t_j} + \left(v_\theta(x_{t_j}, c, t_j) + \frac{\sigma_{t_j}^2}{2t_j}(x_{t_j} + (1-t_j)v_\theta(x_{t_j}, c, t_j)) \right) \Delta t + \sigma_{t_j}\sqrt{-\Delta t}\,\varepsilon_{t_j}, \tag{2}$$

where $v_\theta(x_{t_j}, c, t_j)$ predicts the velocity field, $\varepsilon_{t_j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the injected Gaussian noise, and σ_{t_j} is the coefficient controlling the noise scale. Next, we can express the corresponding Gaussian transition kernel of Eq. (2) as:

$$\begin{aligned} p_\theta(x_{t_j+\Delta t} \mid x_{t_j}, c) &= \mathcal{N}(\mu_\theta(x_{t_j}, c, t_j), -\sigma_{t_j}^2(\Delta t)\mathbf{I}) \\ \mu_\theta(x_{t_j}, c, t_j) &= x_{t_j} + \left(v_\theta(x_{t_j}, c, t_j) + \frac{\sigma_{t_j}^2}{2t_j}(x_{t_j} + (1-t_j)v_\theta(x_{t_j}, c, t_j)) \right) \Delta t. \end{aligned} \tag{3}$$

Based on Eq. (3), we can express the closed-form of KL-divergence in Eq. (1) as:

$$\text{KL}\Big(p^S_\theta(\cdot \mid x_{t_j}, c) \,\Big\|\, p^T_\phi(\cdot \mid x_{t_j}, c) \Big) = \frac{1}{2\sigma_{t_j}^2(-\Delta t)} \left\| \mu^S_\theta(x_{t_j}, c, t_j) - \mu^T_\phi(x_{t_j}, c, t_j) \right\|_2^2, \tag{4}$$

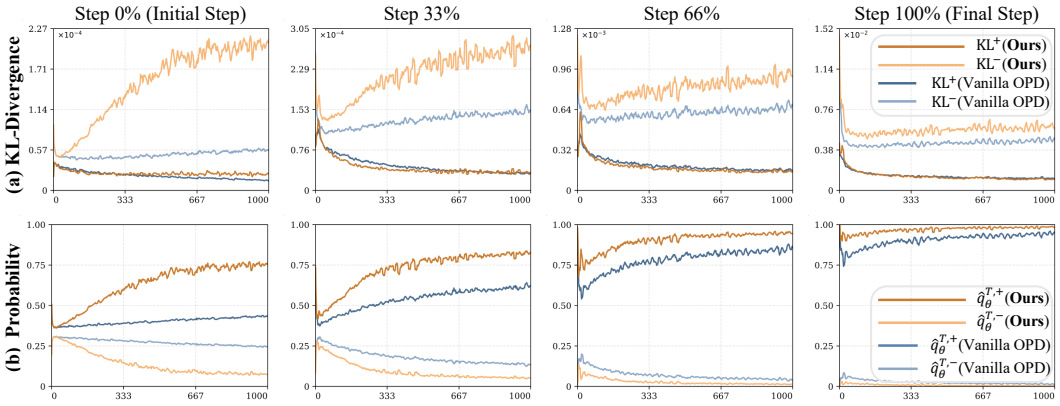


Figure 2: **(a) The KL-divergence between the student and different teachers.** We use KL^+ and KL^- to denote the KL-divergences for task-specific and irrelevant teachers, respectively. Compared with our D-OPD, the numerical discrepancy between KL^+ and KL^- remains relatively narrow under the vanilla OPD objective. **(b) The derived probabilities of the task expert ($\hat{q}_\theta^{T,+}$) and irrelevant teachers ($\hat{q}_\theta^{T,-}$).** In contrast to the vanilla OPD, the disparity between $\hat{q}_\theta^{T,+}$ and $\hat{q}_\theta^{T,-}$ in our method is more pronounced, achieving better inter-teacher discriminability.

where $\mu_\theta^S(x_{t_j}, c, t_j)$ and $\mu_\phi^T(x_{t_j}, c, t_j)$ are computed from $v_\theta^S(x_{t_j}, c, t_j)$ and $v_\phi^T(x_{t_j}, c, t_j)$, respectively. We then formulate the vanilla OPD objective for the multi-task scenario:

$$\mathcal{L}_{\text{OPD}}^{\text{Vanilla}} = \mathbb{E}_{m\sim\pi_m, c_m\sim D_m, x_{t_{T:1}}\sim p_{\theta_{\text{old}}}^S(c_m)} \left[\sum_{j=T}^{1} \mathrm{KL}\Big(p_\theta^S(\cdot \mid x_{t_j}, c_m) \,\Big\|\, p_{\phi_m}^T(\cdot \mid x_{t_j}, c_m) \Big) \right], \tag{5}$$

where the task-index m is sampled from a predefined distribution π_m. Given the prompt c_m sampled from the task dataset D_m, $\mathcal{L}_{\text{OPD}}^{\text{Vanilla}}$ aims to align the student model with the task-specific teacher model $p_{\phi_m}^T$ selected by a hard routing mechanism, which assigns a unique domain expert for each task. Subsequently, the gradient of Eq. (5) can be expressed as:

$$\nabla_\theta \mathcal{L}_{\text{OPD}}^{\text{Vanilla}} = \mathbb{E} \left[\sum_{j=T}^{1} \nabla_\theta \mathrm{KL}\Big(p_\theta^S(\cdot \mid x_{t_j}, c_m) \,\Big\|\, p_{\phi_m}^T(\cdot \mid x_{t_j}, c_m) \Big) \right]. \tag{6}$$

3.2 The Limitations of Vanilla OPD

Before delving into the details of our solution, we first investigate the limitations of vanilla OPD in multi-task scenarios. To this end, we apply the teacher models employed in Flow-OPD Fang et al. (2026) to construct a 3-teacher experiment for OCR Cui et al. (2025), PickScore Kirstain et al. (2023), and GenEval Ghosh et al. (2023) tasks. We plot the values of the KL-divergence between the student and different teacher models across multiple rollout steps in Fig. 2 (a). Specifically, the values corresponding to the task-specific and irrelevant teacher models are denoted as KL^+ and KL^-, respectively. As shown in Fig. 2 (a), the numerical discrepancy between KL^+ and KL^- remains relatively narrow throughout the optimization process under the vanilla OPD objective (Eq. (5)). This indicates that it fails to establish distributional separation between the student and irrelevant teachers, preventing the student from effectively disentangling the extraneous knowledge for each task. To better demonstrate the constraint, we construct a multi-teacher categorical distribution $\hat{q}_\theta^T$ based on the student-teacher discrepancy, which specifies the likelihood of a student's rollout state being attributed to each teacher. In particular, the probability of the n-th teacher is formulated as:

$$\hat{q}_\theta^T(x_{t_j}, c_m)[n] = \frac{\exp\Big(-\mathrm{KL}\big(p_\theta^S(\cdot \mid x_{t_j}, c_m) \,\big\|\, p_{\phi_n}^T(\cdot \mid x_{t_j}, c_m)\big)/\tau_{t_j} \Big)}{\sum_{k=1}^{N_m} \exp\Big(-\mathrm{KL}\big(p_\theta^S(\cdot \mid x_{t_j}, c_m) \,\big\|\, p_{\phi_k}^T(\cdot \mid x_{t_j}, c_m)\big)/\tau_{t_j} \Big)}, \quad n \in [1, N_m], \tag{7}$$

where N_m and τ_{t_j} denote the teacher number and temperature coefficient, respectively. Intuitively, a higher $\hat{q}_\theta^T[n]$ indicates that the student distribution is more closely aligned with the corresponding teacher $p_{\phi_n}^T$ than with others. Next, we report this probability of the task expert ($\hat{q}_\theta^{T,+} = \hat{q}_\theta^T[m]$) and the averaged probabilities of irrelevant teachers ($\hat{q}_\theta^{T,-} = \frac{\sum_{k\neq m}\hat{q}_\theta^T[k]}{(N_m-1)}$) in Fig. 2 (b). As shown in the figure, the disparity between $\hat{q}_\theta^{T,+}$ and $\hat{q}_\theta^{T,-}$ remains limited in vanilla OPD, particularly in high-noise steps that capture global semantics. We posit that this leads to low *inter-teacher discriminability*, indicating that the student's rollouts are indistinguishable across teachers, thereby obscuring which expert the model is currently imitating for a given task. As evident in our experiment, this leads to suboptimal training efficiency and performance ceiling.

3.3 Our Solution: D-OPD

Based on the preceding analysis, we propose **D-OPD**, a simple yet efficient OPD framework designed to address the above limitations. In essence, it reformulates the multi-teacher on-policy distillation process as a task-expert classification problem. To achieve this, D-OPD proposes the following InfoNCE-style Oord et al. (2018) cross-entropy (CE) objective to increase the probability of the task-specific teacher $\hat{q}_\theta^{T,+} = \hat{q}_\theta^T[m]$ established in Eq. (7):

$$\begin{aligned}\mathcal{L}_{\text{CE}} &= \mathbb{E}_{m\sim\pi_m, c_m\sim D_m, x_{t_{T:1}}\sim p_{\theta_{\text{old}}}^S(c_m)}\left[\sum_{j=T}^{1} -\log(\hat{q}_\theta^{T,+})\right] \\ &= \mathbb{E}\left[\sum_{j=T}^{1} -\log\left(\frac{\exp\Big(-\text{KL}\big(p_\theta^S(\cdot\mid x_{t_j}, c_m)\,\Big\|\,p_{\phi_m}^T(\cdot\mid x_{t_j}, c_m)\big)/\tau_{t_j}\Big)}{\sum_{k=1}^{N_m}\exp\Big(-\text{KL}\big(p_\theta^S(\cdot\mid x_{t_j}, c_m)\,\Big\|\,p_{\phi_k}^T(\cdot\mid x_{t_j}, c_m)\big)/\tau_{t_j}\Big)}\right)\right],\end{aligned} \tag{8}$$

Analogous to the InfoNCE loss used in contrastive learning, L_{CE} maximizes the similarity (negative KL-divergence) between the student and the task-specific teacher (positive pair), while minimizing it with other irrelevant ones (negative pairs). Notably, Eq. (8) is formulated under a hard routing mechanism, wherein each c_m is assigned to a unique teacher $p_{\phi_m}^T$. More generally, Eq. (8) can be extended to a soft-label version that aligns the categorical distribution in Eq. (7) with a designated expert-centered distribution $q_{c_m}^T$:

$$\mathcal{L}_{\text{CE}}^{\text{Soft}} = \mathbb{E}_{m\sim\pi_m, c_m\sim D_m, x_{t_{T:1}}\sim p_{\theta_{\text{old}}}^S(c_m)}\left[\sum_{j=T}^{1}\sum_{n=1}^{N_m} -q_{c_m}^T[n]\log(\hat{q}_\theta^T[n])\right], \tag{9}$$

where $q_{c_m}^T$ reflects the desired contribution of each teacher's guidance conditioned on c_m. In practice, constructing $q_{c_m}^T$ requires an external LLM or a rule-based mapping strategy. To demonstrate the effectiveness, we further derive the gradient of L_{CE} and $L_{\text{CE}}^{\text{Soft}}$:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{\text{CE}} &= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}\frac{\mathbb{1}_{n=m}-\hat{q}_\theta^T[n]}{\tau_{t_j}}\nabla_\theta\text{KL}\left(p_\theta^S(\cdot\mid x_{t_j}, c_m)\,\Big\|\,p_{\phi_n}^T(\cdot\mid x_{t_j}, c_m)\right)\right)\right] \\ \nabla_\theta\mathcal{L}_{\text{CE}}^{\text{Soft}} &= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]-\hat{q}_\theta^T[n]}{\tau_{t_j}}\nabla_\theta\text{KL}\left(p_\theta^S(\cdot\mid x_{t_j}, c_m)\,\Big\|\,p_{\phi_n}^T(\cdot\mid x_{t_j}, c_m)\right)\right)\right],\end{aligned} \tag{10}$$

where $\mathbb{1}_{n=m}$ is the indicator function. The comprehensive derivation is provided in the appendix. Eq. (10) exhibits the distinct optimization dynamics between D-OPD and vanilla OPD. As shown in Eq. (6), vanilla OPD merely minimizes the KL-divergence between the student and the task expert selected via the hard routing mechanism. In contrast, $L_{\text{CE}}^{\text{Soft}}$ simultaneously pushes the student's output distribution away from other irrelevant teachers (larger KL-divergence) when the predictive probability exceeds the predefined threshold (*i.e.*, $\hat{q}_\theta^T[n] > q_{c_m}^T[n]$). On the other hand, L_{CE} consistently induces a repulsive force that distances the student from irrelevant teachers. In the following, we adopt Eq. (8) in D-OPD due to its effectiveness and simplicity.

As depicted in Fig. 2 (b), the numerical discrepancy between $\hat{q}_\theta^{T,+}$ and $\hat{q}_\theta^{T,-}$ increases significantly under the proposed D-OPD, leading to better inter-teacher discriminability. Furthermore, a higher

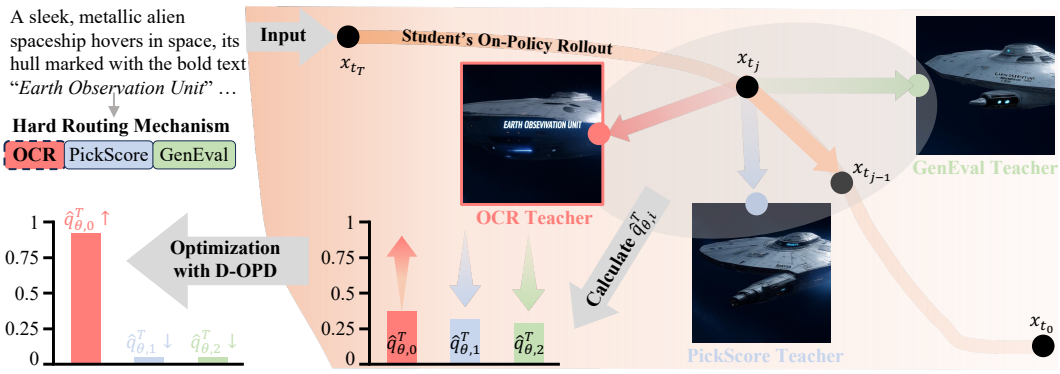


Figure 3: The core framework of the proposed D-OPD. We use different colors to represent the student and teacher models. For this example, the input prompt is sampled from the OCR task. By optimizing the student model with D-OPD, our method pushes the student closer to the task-specific teacher (red color, $\hat{q}_\theta^T \uparrow$) and distances it from irrelevant ones (blue and green colors, $\hat{q}_\theta^T \downarrow$), improving the inter-teacher discriminability of the student's rollout states.

$\hat{q}_\theta^{T,+}$ indicates that the distributional discrepancy between the student and the task-specific teacher is substantially smaller than that with others, thereby facilitating the decoupling of the knowledge from irrelevant teachers. We also present the core framework of our D-OPD in Fig. 3.

Despite its effectiveness, the current D-OPD is hindered by the computational overhead of computing Eq. (7), which entails a growing number of function evaluations (NFEs) as the number of teachers increases. To alleviate this issue, we further introduce the following improvements.

Dual-Teacher Strategy. To avoid the excessive NFEs caused by multiple teachers, we construct a dual-teacher categorical distribution instead of using the full set. For the m-indexed task, one of the teachers is designated as the domain expert $p_{\phi_m}^T$, while the second one is selected as an arbitrary irrelevant teacher $p_{\phi_{m^-}}^T, m^- \neq m$. Then, we can reformulate Eq. (7) as:

$$\begin{aligned}
\hat{q}_\theta^T(x_{t_j}, c_m)[0] &= \frac{\exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_m}^T)/\tau_{t_j}\right)}{\mathrm{sg}\left(\exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_m}^T)/\tau_{t_j}\right)\right) + \exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_{m^-}}^T)/\tau_{t_j}\right)} \\
\hat{q}_\theta^T(x_{t_j}, c_m)[1] &= \frac{\exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_{m^-}}^T)/\tau_{t_j}\right)}{\mathrm{sg}\left(\exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_m}^T)/\tau_{t_j}\right)\right) + \exp\left(-\mathrm{KL}(p_\theta^S \,\|\, p_{\phi_{m^-}}^T)/\tau_{t_j}\right)},
\end{aligned} \tag{11}$$

where we apply the stop-gradient operation $\mathrm{sg}(\cdot)$ to the first term in the denominator to stabilize training. More details are provided in the appendix. Therefore, we can substitute the $\hat{q}_\theta^{T,+}$ term from Eq. (8) with $\hat{q}_\theta^T[0]$ in Eq. (11), enabling more efficient training. We refer to this mechanism as the *Dual-Teacher* (DT) strategy.

Early-Window Strategy. As shown in Fig. 2, the improvement of D-OPD over vanilla OPD is particularly pronounced during early steps and gradually diminishes in subsequent stages. This prompts a critical question: *Is it sufficient to employ Eq.* (8) *exclusively in high-noise rollout steps*? Fig. 4 further demonstrates this hypothesis by generating multiple student rollouts derived from the same Gaussian noise, where the student model is replaced by different teachers at selected steps. As depicted in the figure, such substitutions in the early steps cause significant deviations in the final outcomes, whereas modifications in the later steps result in negligible discrepancies. This also highlights that increasing the inter-teacher discriminability is crucial during the high-noise steps.

Motivated by this, we introduce the *Early-Window* (EW) strategy, which applies Eq. (8) within the initial time window and reverts to the vanilla OPD objective for the rest of the rollout. Consequently,

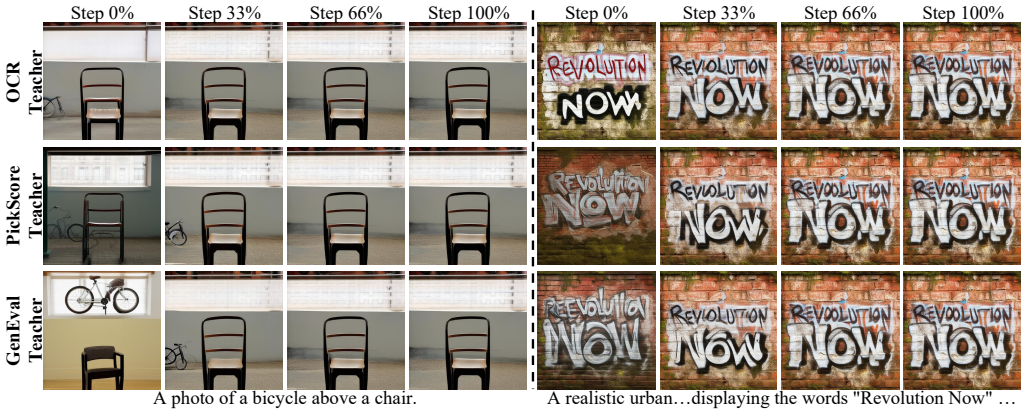


Figure 4: The generated images are sampled from the same Gaussian noise, where the student model is replaced by different teacher models at selected steps. As indicated in the figure, the early substitutions induce a more pronounced impact on the final outcomes, while the replacement in later steps results in marginal discrepancies.

we can derive the final formulation of D-OPD:

$$\mathcal{L}_{\text{D-OPD}} = \begin{cases} \mathcal{L}_{\text{CE}}, & \text{if } t_j \geq t_{\text{trans}}, \\ \mathcal{L}_{\text{OPD}}^{\text{Vanilla}}, & \text{if } t_j < t_{\text{trans}}. \end{cases}, \tag{12}$$

where t_{trans} is the hyperparameter that specifies the window size.

4 Experiment

4.1 Implementation Details

We use Stable Diffusion 3.5 Medium (SD3.5-M) Esser et al. (2024) as the base model. All compared methods are evaluated across the OCR Du et al. (2020), PickScore Kirstain et al. (2023), and GenEval Ghosh et al. (2023) tasks. We use the teacher models and experimental configurations adopted in Flow-OPD Fang et al. (2026). Consequently, we establish 2- and 3-teacher settings depending on the number of jointly optimized tasks. For hyperparameters used in D-OPD, we set τ_{t_j} and t_{trans} to $1e^{-3}$ and $T/3$, respectively. More details and results are provided in the appendix.

For evaluation, we adopt the test splits from Flow-GRPO Liu et al. (2025). Additionally, to assess the generalization capability of the models, we construct a composite benchmark *Multi-EvalBench*. Specifically, we leverage an LLM Wang et al. (2024) to generate 550 text rendering prompts for the OCR task, and use the "Color", "Numeracy", and "Spatial" subcategories from T2I-CompBench++ Huang et al. (2025) to assess the GenEval task. We also report PickScore, ImageReward Xu et al. (2023), Aesthetic, and HPS-v2.1 Wu et al. (2023) scores, evaluating image quality. We compare D-OPD with the following multi-task baselines. **(1) Multi-Reward GRPO (MR-GRPO)** alternately optimizes the GRPO objective Liu et al. (2025) for individual tasks. **(2) Vanilla OPD** adopts the official implementation of Flow-OPD. Since previous studies Li et al. (2026a); Fang et al. (2026) have demonstrated that some alternative paradigms (*e.g.*, Cascade RL, Mixed-Reward RL) are inefficient in multi-task scenarios, they are excluded from our comparisons. The discussion of these methods is provided in the appendix.

4.2 Comparisons with Multi-Task Baselines

In this section, we conduct a comprehensive comparison between our method and the multi-task baselines. As shown in Fig. 5, D-OPD consistently outperforms other approaches in convergence speed across all settings, where our method achieves high rewards in very few training iterations. In particular, Fig. 5 (b) depicts the per-task reward curve under the 3-teacher setting, demonstrating that D-OPD improves the training efficiency for each independent task.

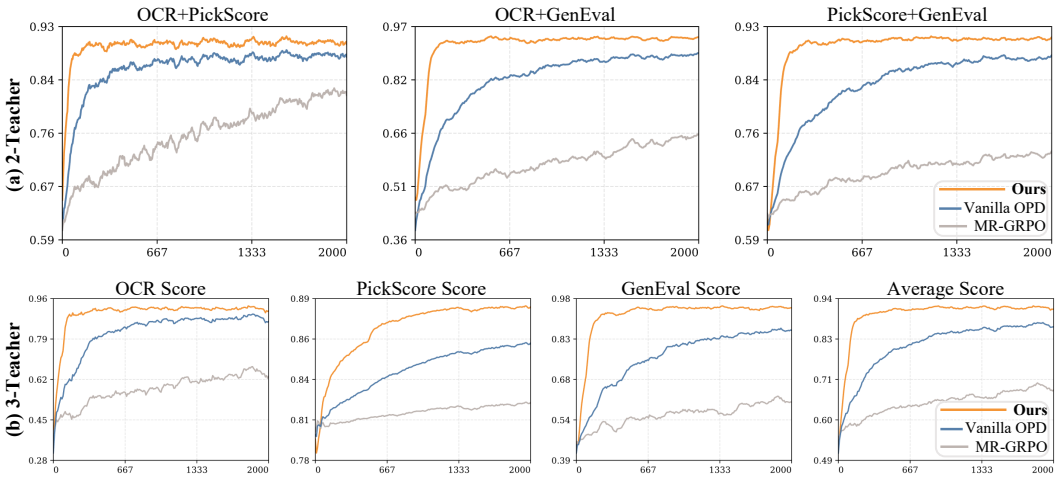


Figure 5: The training reward curves demonstrate that our method achieves significantly faster convergence than both MR-GRPO and vanilla OPD in 2- and 3-teacher settings.

Table 1: Evaluation on test splits of OCR, PickScore, and GenEval tasks.

Model	OCR	PickScore	GenEval	Average Score
SD3.5-M	0.57	21.70	0.61	0.6715
OCR-Teacher	0.93	21.73	0.62	0.7952
PickScore-Teacher	0.71	23.25	0.46	0.6881
GenEval-Teacher	0.64	21.60	0.94	0.8036
MR-GRPO	0.74	21.92	0.74	0.7744
Vanilla OPD	0.90	22.80	0.89	0.8889
D-OPD w/o DT, EW	**0.94**	23.17	0.94	0.9237
D-OPD w/o EW	**0.94**	23.14	0.94	0.9233
D-OPD (ours)	**0.94**	**23.31**	**0.96**	**0.9321**

Table 2: Comparison on MultiEvalBench.

Model	OCR	Color	Numeracy	Spatial	PickScore	Aesthetic	ImageReward	HPSv2.1
SD3.5-M	0.55	0.79	0.63	0.29	22.52	5.49	1.17	0.2944
MR-GRPO	0.71	0.80	0.66	0.34	22.74	5.51	1.30	0.2990
Vanilla OPD	0.87	0.83	0.72	0.46	23.06	5.68	1.50	0.3062
D-OPD (ours)	**0.94**	**0.84**	**0.73**	**0.49**	**23.96**	**6.35**	**1.60**	**0.3337**

Tab. 1 reports the performance of the compared methods in the test split of each task. Despite improving overall performance over the pre-trained model, MR-GRPO and vanilla OPD are limited by suboptimal training efficiency and performance ceiling. In contrast, D-OPD achieves the best metric for all tasks. Tab. 2 presents the results for the proposed MultiEvalBench, highlighting the better generalization capability and image quality of our method. In addition, Fig. 6 shows qualitative results. As shown in the figure, the images generated by D-OPD exhibit better text-image alignment and aesthetic quality compared to other baselines.

4.3 Ablation Studies

In this section, we investigate the impact of each proposed component. First, we assess the effectiveness of the cross-entropy objective (w/o DT, EW). As indicated by Tab. 1 and the reward curves in Fig. 7, the current model still outperforms both MR-GRPO and vanilla OPD in convergence speed and task metrics. Furthermore, incorporating DT (w/o EW) yields comparable training efficiency and performance to the basic setting (w/o DT, EW), while reducing computational over-

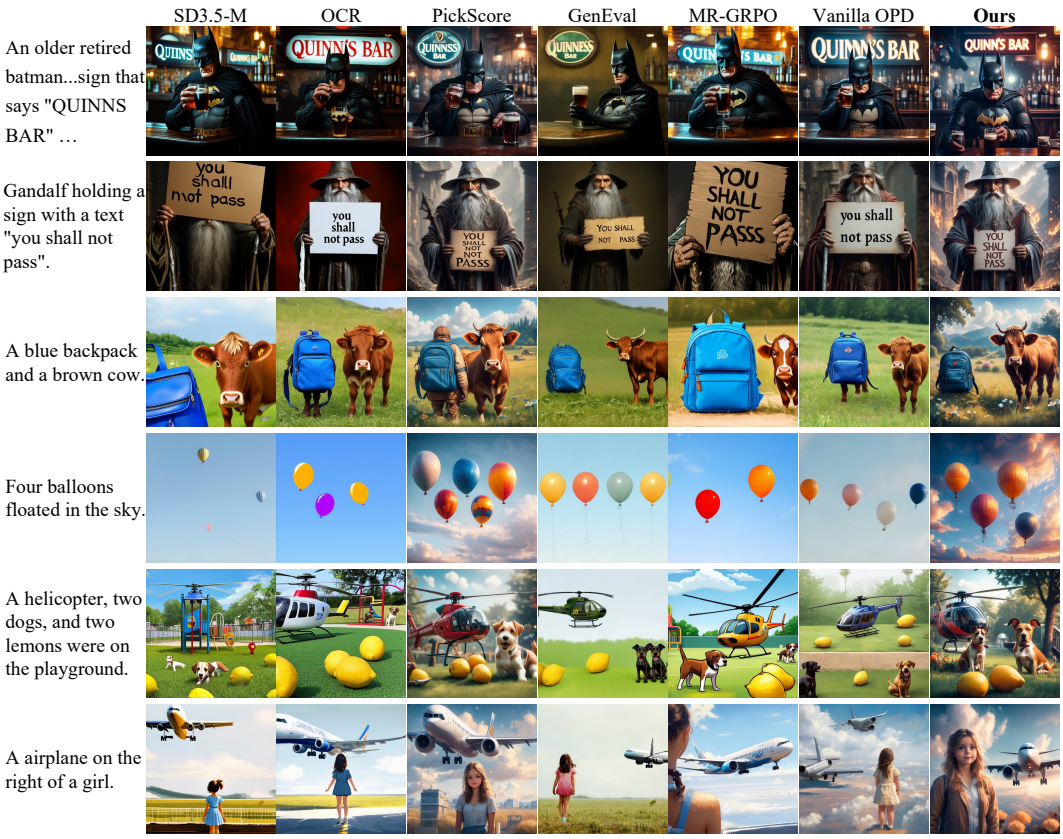


Figure 6: Qualitative results. As shown in the figures, the images generated from D-OPD outperform others in text-image alignment and aesthetic quality.

head incurred by multiple teachers. Additionally, Fig. 7 illustrates the significant improvement in convergence speed after introducing the EW mechanism (**Ours**).

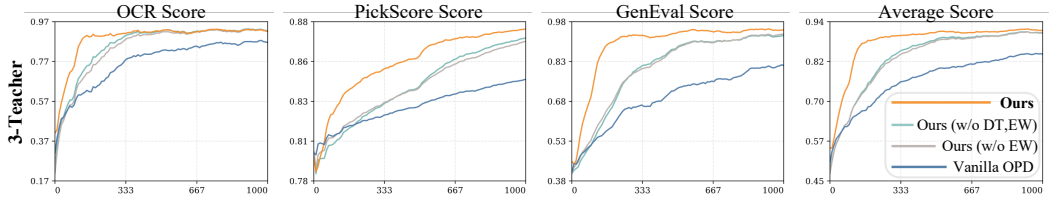


Figure 7: Ablation studies. As shown in the figures, incorporating DT (w/o EW) achieves a comparable convergence speed to the basic setting (w/o DT,EW), while reducing the computational overhead. In addition, the introduction of EW (**Ours**) enhances the training efficiency.

5 Conclusion

In this paper, we introduce D-OPD, an efficient framework for improving OPD in multi-task settings. By reformulating multi-teacher on-policy distillation as task-expert classification, D-OPD aligns the student with the task-specific teacher while separating it from irrelevant ones, achieving better inter-teacher discriminability of student rollouts than vanilla OPD. It further incorporates DT and EW to reduce multi-teacher computational overhead and enhance training efficiency. Extensive experiments show that D-OPD significantly outperforms existing RL and OPD baselines in overall performance, demonstrating its strong potential for multi-task post-training.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the readability, grammar, and linguistic clarity of the manuscript. Generative AI tools were not used to generate research ideas, formulate hypotheses or mathematical claims, develop theoretical models, design the methodology or experiments, implement the proposed method, analyze data, interpret results, or draw scientific conclusions. All AI-assisted edits were carefully reviewed and verified by the authors, and no scientific content or claims were introduced without independent author validation. We take full responsibility for the final content of this work.

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 *ICLR*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv*, 2020.

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 *ICLR*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 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*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *ICLR*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *TPAMI*, 2025.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv*, 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 *NeurIPS*, 2023.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv*, 2025.

Black Forest Labs. Flux, 2024.

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*, 2026a.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv*, 2025.

Kevin Lu and Thinking Machines Lab. On-Policy Distillation, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv*, 2017.

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 *CVPR*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024.

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*, 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 *NeurIPS*, 2023.

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

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *ACL*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv*, 2026.

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 *ICLR*, 2025.

A Appendix

A.1 Multi-Task Post-Training Paradigms

In the following, we list some common multi-task post-training algorithms in T2I generation. Most of them have been discussed in the previous literature Fang et al. (2026); Li et al. (2026a).

(1). Mixed-Reward RL combines the rewards from multiple tasks to derive a single value via a predefined weighting scheme. Limited by a sparse information bottleneck, it prefers to improve simpler tasks during optimization, leading to degraded performance in more challenging ones.

(2). Cascade RL sequentially optimizes across different tasks, employing a single-reward RL method during each stage, such as GRPO Liu et al. (2025) or DiffusionNFT Zheng et al. (2025). Although it circumvents the gradient conflicts inherent in simultaneous training, it is severely hindered by catastrophic forgetting, resulting in performance degradation in previously learned tasks. As reported in Diffusion-OPD Li et al. (2026a), such a paradigm suffers from slow convergence speed.

(3). Multi-Task RL alternately optimizes individual tasks, thereby mitigating the catastrophic forgetting issue in Cascade RL. However, constrained by the sparse scalar reward, it exacerbates gradient conflicts among distinct objectives, thereby impairing training efficiency. Notably, the Multi-Reward GRPO (MR-GRPO) discussed in our main paper belongs to this category.

A.2 More Details of D-OPD Method

Derivation of Eq. (10). We first derive the gradient of the soft-label cross-entropy objective $L_{\text{CE}}^{\text{Soft}}$:

$$
\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{CE}}^{\text{Soft}} &= \mathbb{E}\left[\sum_{j=T}^{1}\sum_{n=1}^{N_m}\nabla_\theta\Big(-q_{c_m}^T[n]\log\left(\hat{q}_\theta^T[n]\right)\Big)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\sum_{n=1}^{N_m}\left(-q_{c_m}^T[n]\nabla_\theta\log\left(\hat{q}_\theta^T[n]\right)\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(-\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]}{\hat{q}_\theta^T[n]}\nabla_\theta\hat{q}_\theta^T[n]\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(-\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]}{\hat{q}_\theta^T[n]}\sum_{k=1}^{N_m}\frac{\partial\hat{q}_\theta^T[n]}{\partial(-\mathrm{KL}_k/\tau_{t_j})}(-\nabla_\theta\mathrm{KL}_k/\tau_{t_j})\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]}{\hat{q}_\theta^T[n]}\sum_{k=1}^{N_m}\hat{q}_\theta^T[n](\mathbb{1}_{n=k}-\hat{q}_\theta^T[k])\nabla_\theta\mathrm{KL}_k/\tau_{t_j}\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}q_{c_m}^T[n]\sum_{k=1}^{N_m}(\mathbb{1}_{n=k}-\hat{q}_\theta^T[k])\nabla_\theta\mathrm{KL}_k/\tau_{t_j}\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}q_{c_m}^T[n]\nabla_\theta\mathrm{KL}_n/\tau_{t_j}-\sum_{k=1}^{N_m}\hat{q}_\theta^T[k]\nabla_\theta\mathrm{KL}_k/\tau_{t_j}\right)\right] \\
&= \mathbb{E}\left[\sum_{j=T}^{1}\left(\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]-\hat{q}_\theta^T[n]}{\tau_{t_j}}\nabla_\theta\mathrm{KL}\left(p_\theta^S(\cdot\mid x_{t_j},c_m)\,\middle\|\,p_{\phi_n}^T(\cdot\mid x_{t_j},c_m)\right)\right)\right].
\end{aligned}
\tag{13}
$$

We can further derive the $\nabla_\theta L_{\text{CE}}$ in Eq. (10) by setting $q_{c_m}^T[n]$ to a one-hot vector (*i.e.*, $q_{c_m}^T[m]=1$ with zeros elsewhere). The weighting coefficient of the gradient associated with the n-index teacher can be expressed as:

$$\frac{\mathcal{L}_{\text{CE}}^{\text{Soft}}}{\partial \text{KL}_n} = \mathbb{E}\left[\sum_{j=T}^{1}\left(\frac{q_{c_m}^T[n] - \hat{q}_\theta^T[n]}{\tau_{t_j}}\right)\right]. \tag{14}$$

The stop gradient operation used in Eq. (11). As indicated in Fig. 2 from the main paper, the multi-teacher categorical distributions in high-noise steps are nearly uniform at the beginning of training. However, the large $\hat{q}_\theta^T[n]$ diminishes the gradient weight, as evidenced in Eq. (14), thereby impeding the student's learning from the task expert and resulting in training instability. This is particularly pronounced when the number of teachers is small, such as the 2-teacher setup adopted in the proposed Dual-Teacher (DT) strategy. Therefore, we apply the stop-gradient operation to the first term in the denominator of Eq. (11), resulting in the following gradient associated with the task-specific teacher $p_{\phi_m}^T$:

$$\begin{aligned}\frac{\mathcal{L}_{\text{CE}}^{\text{Soft}}}{\partial \text{KL}_m} &= \mathbb{E}\left[\sum_{j=T}^{1}\left(-\sum_{n=1}^{N_m}\frac{q_{c_m}^T[n]}{\hat{q}_\theta^T[n]}\frac{\partial \hat{q}_\theta^T[n]}{\partial(-\text{KL}_m/\tau_{t_j})}(-1/\tau_{t_j})\right)\right] \\ &= \mathbb{E}\left[\sum_{j=T}^{1}\frac{q_{c_m}^T[m]}{\hat{q}_\theta^T[m]}\hat{q}_\theta^T[m](1/\tau_{t_j})\right] \\ &= \mathbb{E}\left[\sum_{j=T}^{1}\frac{q_{c_m}^T[m]}{\tau_{t_j}}\right].\end{aligned} \tag{15}$$

Unlike Eq. (14), Eq. (15) amplifies the gradient weight for the m-indexed teacher, while keeping the weights of the remaining teachers unchanged.

The training pipeline of D-OPD. The overall training pipeline of our D-OPD is summarized in Algorithm 1.

Algorithm 1 The overall training pipeline of D-OPD.

Input: Task datasets $\{D_n\}_{n=1}^{N_m}$; teacher models $\left\{p_{\phi_n}^T\right\}_{n=1}^{N_m}$; student model p_θ^S; early window size hyperparameter t_{trans}; distribution for sampling the task index π_m; total step T;
Output: The optimized parameters of the student model θ.

while True **do**
  Sample the task index m from π_m and get the corresponding dataset D_m;
  Sample prompts c_m from D_m;
  Sample x_{t_T} from Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$;
  $L_{\text{total}} = 0$;
  for all j in $[T, ..., 1]$ **do**
    if $t_j \geq t_{\text{trans}}$ **then** ▷ Applying Eq. (8)
      Randomly select an irrelevant teacher $p_{\phi_{m^-}}^T$, where $m^- \neq m$;
      Calculate $\hat{q}_\theta^T[0]$ using Eq. (11);
      $L_{\text{step}} = -\log(\hat{q}_\theta^T[0])$; ▷ Using hard routing mechanism
    else ▷ Applying Eq. (5)
      $L_{\text{step}} = \text{KL}\Big(p_\theta^S(\cdot \mid x_{t_j}, c_m) \,\Big\|\, p_{\phi_m}^T(\cdot \mid x_{t_j}, c_m)\Big)$;
    end if
    $L_{\text{total}} += L_{\text{step}}$;
    Derive $x_{t_{j-1}}$ using Eq. (2);
  end for
  Update θ using $\nabla_\theta L_{\text{total}}$;
end while

Table 3: Evaluation on test splits of OCR, PickScore, GenEval, and customization tasks.

Model	OCR	PickScore	GenEval	DINOv2-1	DINOv2-2	Average Score
SD3.5-M	0.57	21.70	0.61	0.68	0.62	0.6589
Vanilla OPD	<u>0.89</u>	<u>22.58</u>	<u>0.74</u>	<u>0.87</u>	<u>0.77</u>	<u>0.8277</u>
D-OPD (ours)	**0.93**	**23.22**	**0.94**	**0.89**	**0.78**	**0.8866**

A.3 More Results

Implementation Details. D-OPD is built upon the official codebase of Flow-OPD Fang et al. (2026), adopting its provided teacher models and training configurations. For the implementation of the vanilla OPD, we omit some technical tricks used in Flow-OPD, such as "Merging-as-Initialization" and "Manifold Anchor Regularization", isolating the effect of the OPD objective. Specifically, the image resolution is set to 512×512, with a group size of 24, and a learning rate of $1e^{-4}$. For the SDE scheduler, we use 10 sampling steps, a noise level of 0.7, and a classifier-free guidance Ho & Salimans (2022) (CFG) scale of 4.5. The LoRA rank is set to 32. Furthermore, to prevent overfitting, we incorporate a KL divergence loss between the student and pre-trained model for all methods, with a weighting factor of 0.04. In the 3-teacher setting, we adopt a data mixing ratio of $1:1:3$ for OCR Du et al. (2020), PickScore Kirstain et al. (2023), and GenEval Ghosh et al. (2023) tasks. For 2-teacher settings, we apply $1:1$ ratio. For evaluation, we employ 28 sampling steps and a CFG scale of 4.5 to sample 512×512 images. For the proposed MultiEvalBench, to mitigate the counting discrepancy between training and testing inputs, we filter out prompts containing an object number greater than four in the "Numeracy" subcategory.

More Qualitative Results on Test Splits. We present additional qualitative results on the test splits of OCR, PickScore, and GenEval tasks in Fig. 8-Fig. 10. As evident in these figures, our method achieves the best performance for all tasks, demonstrating superior text-image alignment.

Ablation Studies. Furthermore, we investigate the impacts of different hyperparameter settings under the 3-teacher setting. As illustrated in Fig. 11, a smaller temperature coefficient τ accelerates model convergence, peaking at $1e-3$. Moreover, we observe that further decreasing τ induces training instability (*e.g.*, $1e-4$). Consequently, we set τ to $1e-3$. Notably, across all configurations, our method consistently outperforms the vanilla OPD baseline.

Subsequently, Fig. 12 presents the results of employing the cross-entropy (CE) loss at different rollout stages. Applying the CE loss during the final $T/3$ stage yields the slowest convergence, whereas employing it during the middle $T/3$ stage results in suboptimal performance, highlighting the critical importance of enhancing inter-teacher discriminability at earlier steps. Therefore, our Early-Window (EW) strategy applies the CE loss during the initial $T/3$ stage.

More Experimental Settings. We further extend our evaluation to a 5-teacher setting. Beyond the OCR, PickScore, and GenEval tasks, we incorporate two additional teachers tailored for customization tasks, which are obtained by fine-tuning SD3.5-M Esser et al. (2024) on two instance subsets of the DreamBooth dataset Ruiz et al. (2023). In particular, a powerful LLM Wang et al. (2024) is used to generate 8885 and 500 prompts for training and evaluation, respectively. To evaluate the customization task, we calculate the cosine similarity between the DINOv2 Oquab et al. (2023) features derived from the generated images and instance images. As shown in Fig. 13, our method yields significant performance gains over the vanilla OPD baseline. Furthermore, Tab. 3 details the final results of compared methods on the test splits for each task, demonstrating that D-OPD achieves a higher performance ceiling.

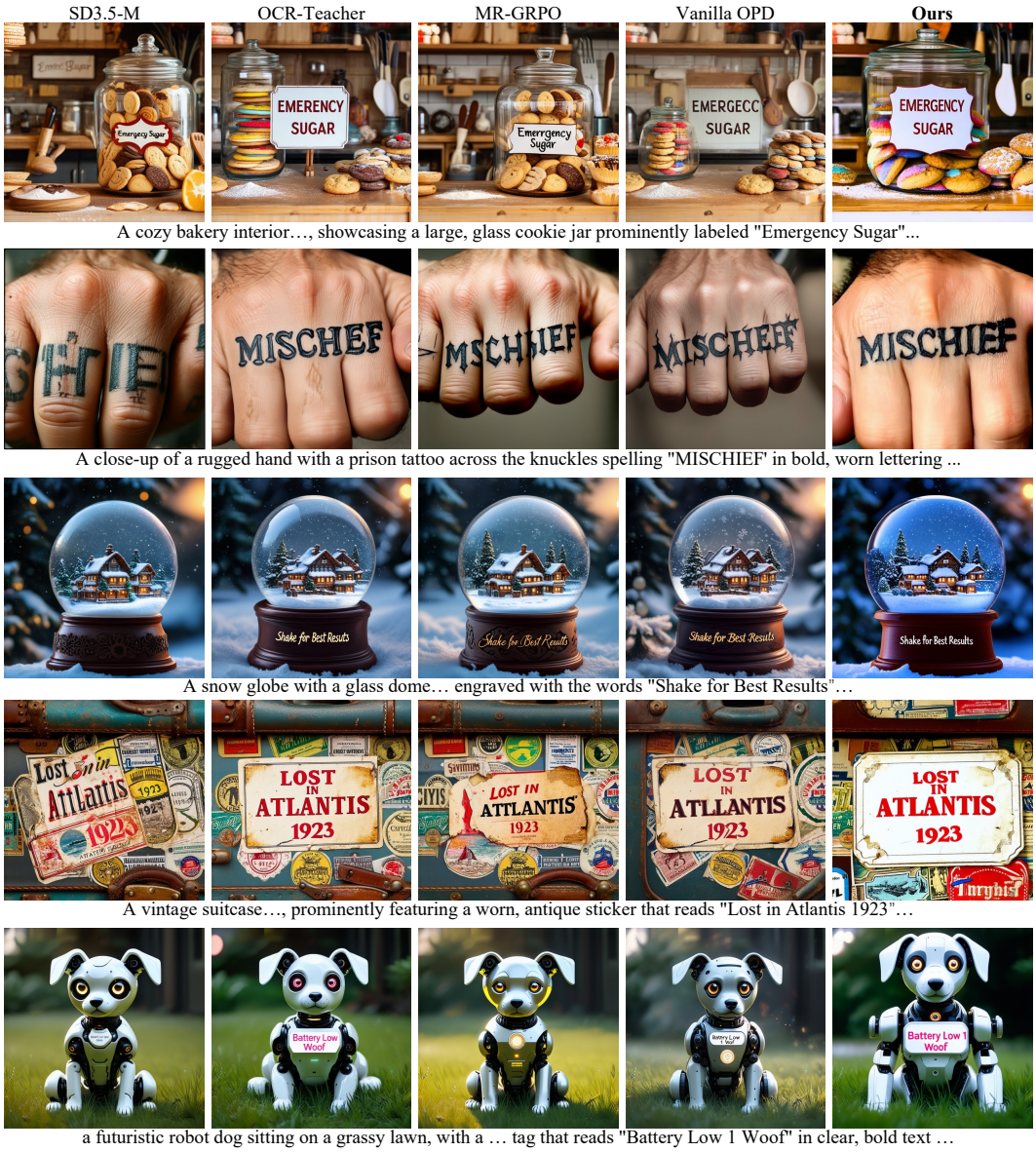


Figure 8: Qualitative results on the test split of OCR Du et al. (2020) task.



Figure 9: Qualitative results on the test split of PickScore Kirstain et al. (2023) task.



Figure 10: Qualitative results on the test split of GenEval Ghosh et al. (2023) task.

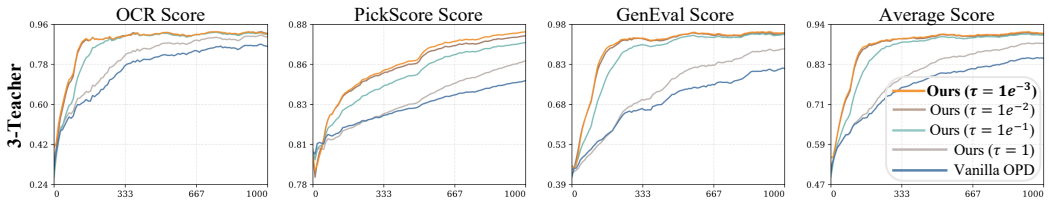


Figure 11: Ablation studies on the temperature τ.

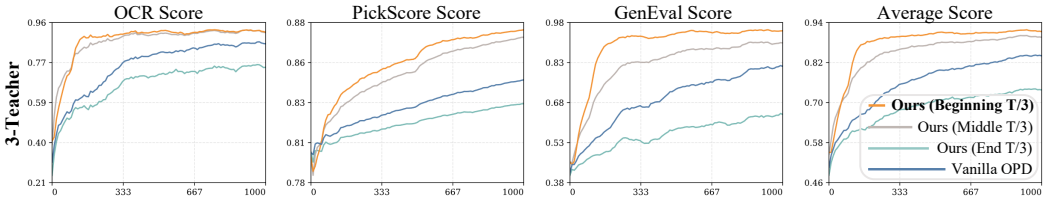


Figure 12: Ablation studies on the application stages of the cross-entropy loss.

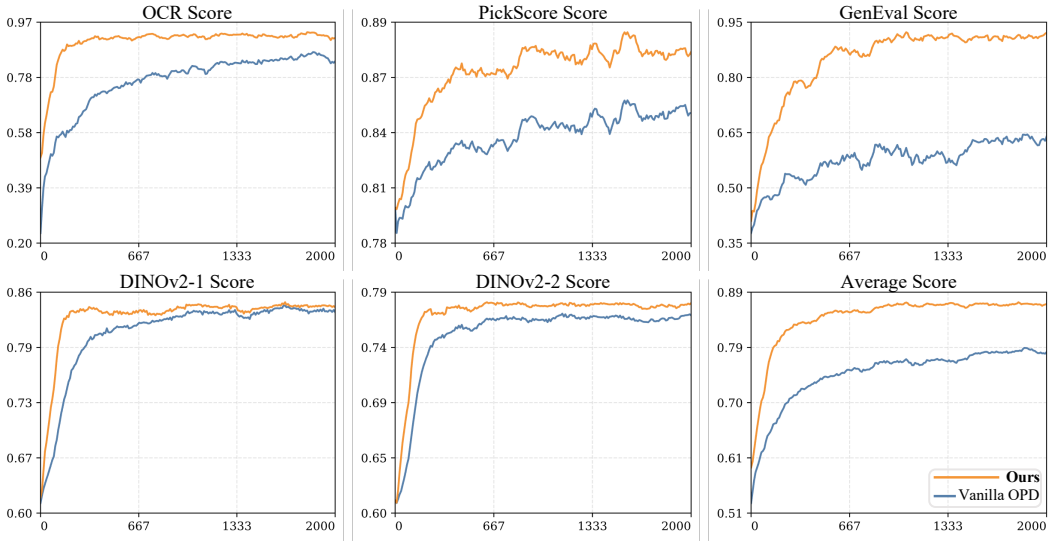


Figure 13: Training reward curves for the 5-teacher setting.