ONE-STAGE FEW-STEP DISTILLATION WITH ON-POLICY TRAJECTORY ANCHORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-aligned diffusion models achieve strong generation performance but typically require many sampling steps. Developing few-step diffusion models that retain strong alignment with human preferences remains an important research goal. On-policy distillation (OPD) transfers teacher-aligned capabilities through local supervision, but its performance degrades sharply with fewer sampling steps. Distribution Matching Distillation (DMD) enables few-step generation but can introduce distributional drift and weaken aligned capabilities. Separating the transfer of teacher-aligned capabilities and few-step distillation into two stages can cause the second stage to undermine the aligned capabilities or few-step generation ability learned in the first. To address these limitations, we propose On-Policy Distribution Matching Distillation (OPDMD), a single-stage framework that jointly enforces teacher anchoring and distribution matching along student-generated trajectories. OPDMD anchors the student velocity to the teacher at each selected student state and applies a distribution-matching correction at the corresponding successor state. We unify these two supervision signals into a transition-aware MSE objective that scales the correction by the transition span. Experiments on SD-3.5-Medium with GRPO-aligned expert teachers show that OPDMD better preserves aligned capabilities and distributional fidelity than OPD, DMD, and their sequential combinations. Additional analyses show that transition-aware correction is especially beneficial under sparse sampling and that balancing the two supervision signals is important for overall performance. These results establish OPDMD as an effective approach to preference-preserving few-step generation.

1 INTRODUCTION

Generative diffusion models (Ho et al., 2020; Lipman et al., 2022; Rombach et al., 2022; Peebles & Xie, 2023; Gu et al., 2025; Wan et al., 2025) have become a leading path to high-fidelity visual synthesis. Beyond visual quality, recent preference alignment methods (Wallace et al., 2024; Liu et al., 2025; Zheng et al., 2026) improve these models with stronger alignment to human preferences. However, these advances typically rely on many sampling steps at inference time, resulting in substantial computational costs that limit practical deployment. Therefore, developing few-step generative models that retain strong alignment with human preferences remains an important research goal (Jiang et al., 2026b).

On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2023; Li et al., 2026a; Fang et al., 2026; Jiang et al., 2026a; Zhou et al., 2026) trains the student with teacher supervision on student-generated states to transfer the teacher's aligned capabilities. In language modeling, OPD (Agarwal et al., 2024; Gu et al., 2023) queries the teacher on states generated by the student, mitigating the exposure mismatch in conventional offline distillation (Bengio et al., 2015). Recent works (Li et al., 2026a; Fang et al., 2026; Zhou et al., 2026; Jiang et al., 2026a) extend this paradigm to diffusion and flow-based generation. These methods match local teacher velocity or transitions along student rollouts to transfer or compose aligned capabilities (Fang et al., 2026; Zhou et al., 2026). Compared with reinforcement learning that depends on scalar rewards (Liu et al., 2025; Zheng et al., 2026), OPD provides dense supervision along student-generated trajectories, offering direct guidance for learning the teacher's preference-aligned behavior.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, OPD still faces challenges in few-step generation. When the student matches the teacher's velocities only on a sparse sampling trajectory, the teacher cannot provide sufficiently accurate guidance for few-step generation. As shown in Fig. 1(a), the aligned capabilities preserved by OPD deteriorate sharply as the number of sampling steps in the training trajectories decreases. A direct remedy is to further distill the OPD-trained student using few-step distillation methods such as Distribution Matching Distillation (DMD) (Yin et al., 2024b;a). DMD provides a global correction signal by minimizing an approximate KL divergence between the student and teacher distributions and has been widely adopted in visual generation (Seedream et al., 2025; Cai et al., 2025). However, we find that applying OPD and DMD in either order introduces a structural mismatch between their objectives. In the OPD-then-DMD pipeline, OPD first anchors the student to the teacher's local behavior, but the subsequent DMD stage can cause the student's distribution to drift away from the teacher's. This drift can weaken the student's aligned capabilities and alter global visual statistics, resulting in overly saturated, high-contrast outputs (Poole et al., 2022; Chen et al., 2026), as illustrated in Fig. 1(b). DMD-then-OPD does not resolve this mismatch: DMD first trains the student for few-step generation. The subsequent OPD stage, limited by inaccurate teacher supervision on a sparse trajectory, undermines the student's learned few-step generation capabilities, leading to severe quality degradation and blurred outputs, as shown in Fig. 1(c, d). These observations suggest that local on-policy teacher anchoring and global distribution correction should be enforced jointly throughout training, rather than applied as two separate stages.

To address this issue, we propose **On-Policy Distribution Matching Distillation** (OPDMD), a single-stage framework that jointly preserves teacher-aligned behavior and corrects the few-step student distribution. The student first generates a few-step trajectory. At each selected source state on this trajectory, we query the teacher to obtain a local velocity target, anchoring the student to the teacher's behavior at the state. Since such local supervision alone cannot compensate for the error induced by sparse trajectory matching, we construct a successor query state through a one-step student transition and evaluate a distribution-matching correction at this state using the same teacher and an online fake-score model. Since the successor state is induced by the student velocity, mapping its distribution-matching correction back to the source-state velocity naturally introduces the transition span. Based on this transition geometry, we derive a transition-aware objective that coordinates local teacher anchoring and distribution correction throughout training.

We evaluate OPDMD on SD-3.5-Medium (Esser et al., 2024) and its GRPO-aligned expert teachers (Liu et al., 2025) under 4-step and 8-step generation settings. OPDMD consistently outperforms OPD, DMD, and their sequential combinations on OCR accuracy, GenEval, and PickScore, while generating samples closer to the teacher distribution. The gains are particularly pronounced in the challenging 4-step setting, and results on T2I-CompBench++ (Huang et al., 2025) further confirm its effectiveness in preserving compositional text-image alignment. Overall, these results demonstrate that jointly enforcing teacher anchoring and distribution matching provides an effective solution for capability-preserving few-step distillation.

In summary, our main contributions are as follows:

- We propose On-Policy Distribution Matching Distillation, a single-stage framework that jointly applies on-policy teacher anchoring and distribution matching. It enables the preservation of aligned capabilities under few-step generation.
- To coordinate these two goals, we derive a transition-aware objective that combines local teacher anchoring and distribution correction. The objective couples supervision at source and successor states through the student's sampling transition.
- Extensive experiments across text rendering, compositional generation, and human preference alignment demonstrate capability preservation and distributional fidelity using GRPO-aligned SD3.5-Medium teachers at 4 and 8 steps.

2 Related Work

2.1 On-Policy Distillation

On-policy distillation was originally developed for autoregressive language models to reduce the mismatch between the training distribution and the states visited during free generation. By super-

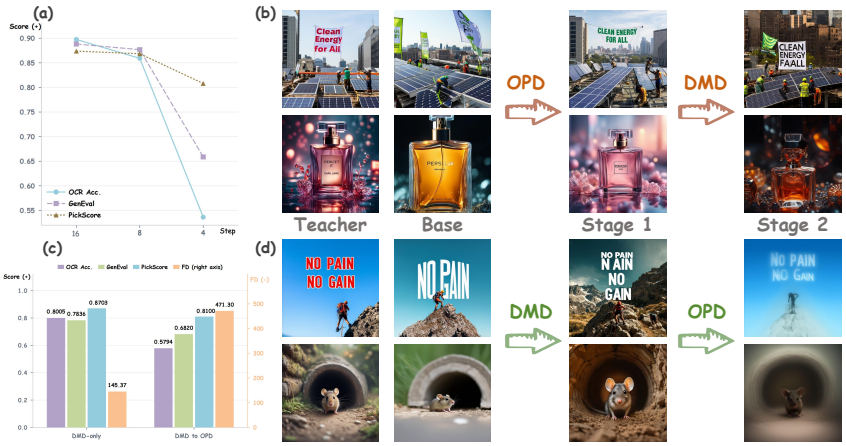

Figure 1: Limitations of OPD and sequential OPD–DMD distillation. (a) OPD exhibits substantial capability degradation as the training trajectories become sparser. (b) OPD-then-DMD improves few-step image quality but introduces noticeable visual distribution shifts. (c–d) DMD-then-OPD further degrades aligned capabilities, increases the feature distance from the teacher distribution, and produces blurry images with missing details.

vising the student on its own generated states, this approach helps address errors that arise during generation (Li et al., 2026b). GKD (Agarwal et al., 2024) queries the teacher on student-generated prefixes, thereby exposing the student to supervision on its own generation distribution. MiniLLM (Gu et al., 2023) optimizes reverse KL divergence between student and teacher sequence distributions via on-policy policy-gradient estimation, leveraging its mode-seeking property to better distill complex teacher distributions into capacity-limited students. Recent works extend the on-policy paradigm to diffusion and flow-based generative models. D-OPSD (Jiang et al., 2026a) improves few-step diffusion models by matching student predictions to richer multimodal self-teacher outputs on student-induced trajectories. DiffusionOPD (Li et al., 2026a) uses per-step KL objectives for diffusion transitions and enables the distillation of multiple reward-specialized teachers into a single student. Flow-OPD (Fang et al., 2026) similarly performs expert-routed supervision along on-policy flow trajectories. DanceOPD (Zhou et al., 2026) formulates source capabilities as velocity fields and composes them through hard-routed velocity matching on student-induced states. These methods demonstrate the effectiveness of on-policy supervision for preserving teacher behavior along the student's generation trajectories. However, their primary focus is continual adaptation or capability composition rather than correcting the distributional error introduced by aggressive step reduction.

2.2 Distribution Matching Distillation

Distribution matching distillation (DMD) (Yin et al., 2024b;a) trains a few-step student by directly matching its generated distribution to the teacher distribution, typically using the difference between teacher and student score estimates as the optimization signal. DMD (Yin et al., 2024b) minimizes an approximate reverse KL divergence between the student and teacher image distributions, using the difference between a frozen teacher score and an online fake score to estimate the distribution-matching direction. It further introduces paired LPIPS (Zhang et al., 2018) regression to stabilize training and mitigate mode dropping. DMD2 (Yin et al., 2024a) removes this paired regression by adopting a two-time-scale update rule that improves the tracking accuracy of the fake-score model, and further incorporates adversarial supervision (Goodfellow et al., 2014) from real data. To support few-step generation, it introduces backward simulation, which exposes the student during training to intermediate states resembling those encountered at inference time. TDM (Luo et al., 2025) extends distribution matching along the student sampling trajectory by aligning intermediate student distributions with the corresponding teacher distributions, and introduces a sampling-steps-aware objective to support flexible sampling steps. These approaches establish distribution matching as an effective framework for training high-quality few-step generators by directly correcting the student distribution. However, distribution-level alignment alone does not explicitly preserve the local transport behavior of a preference-aligned teacher along the evolving student trajectory.

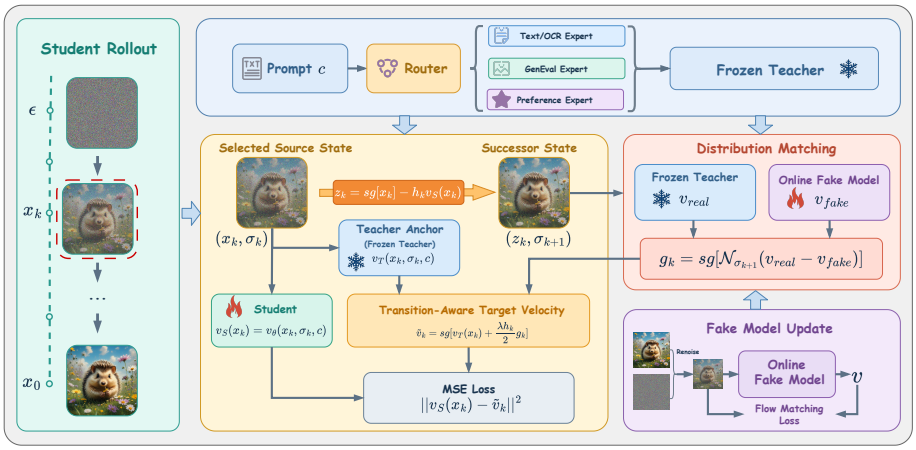

Figure 2: Overview of OPDMD. Along a student-generated few-step trajectory, the teacher provides a local velocity anchor at a selected source state and, together with the online fake model, a distribution-matching correction at the successor state. Both signals form a transition-aware target. The fake model is trained on detached student samples using flow matching. All teacher queries for a given prompt use the same frozen expert selected by the prompt router.

3 METHODS

3.1 PRELIMINARIES

Distribution Matching Distillation (DMD) aligns the student distribution with a pretrained teacher distribution by minimizing the approximate reverse Kullback-Leibler divergence (Yin et al., 2024b):

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_\sigma \left[D_{\text{KL}} \left(p_{S,\sigma} \,\|\, p_{T,\sigma} \right) \right] \tag{1}$$

where $p_{S,\sigma}$ and $p_{T,\sigma}$ denote the noisy student and teacher distributions at noise level σ. Since the score of the student distribution is unavailable, DMD trains an auxiliary fake-score model to estimate the student score from detached student-generated samples. The difference between the frozen teacher score and the learned fake score provides an estimate of the distribution-matching direction (Yin et al., 2024b;a). We parameterize the student, teacher, and fake models with conditional flow-matching velocity fields. For a clean sample $\mathbf{x}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we use the linear interpolation path

$$\mathbf{x}_\sigma = (1 - \sigma)\mathbf{x}_0 + \sigma\boldsymbol{\epsilon} \tag{2}$$

For a sample pair $(\mathbf{x}_0, \boldsymbol{\epsilon})$, the corresponding flow-matching target is $\boldsymbol{\epsilon} - \mathbf{x}_0$. Let $\mathbf{v}_\sigma(\mathbf{x}_\sigma, \mathbf{c})$ denote the conditional velocity field. Under this parameterization, the conditional score and velocity field satisfy the following relation (Lipman et al., 2022) for $\sigma > 0$:

$$\mathbf{s}_\sigma(\mathbf{x}_\sigma, \mathbf{c}) = -\frac{1}{\sigma} \left[\mathbf{x}_\sigma + (1 - \sigma)\mathbf{v}_\sigma(\mathbf{x}_\sigma, \mathbf{c}) \right], \quad \sigma > 0 \tag{3}$$

The corresponding denoised prediction is

$$\widehat{\mathbf{x}}_0 = \mathbf{x}_\sigma - \sigma\mathbf{v}_\sigma(\mathbf{x}_\sigma, \mathbf{c}) \tag{4}$$

For clarity, we use the following notation throughout the method. The student velocity is v_θ, the online fake velocity is v_ϕ, and v_T denotes the velocity of the routed frozen teacher. The operator $\text{sg}[\cdot]$ indicates stop-gradient. We use a decreasing noise grid

$$1 = \sigma_0 > \sigma_1 > \cdots > \sigma_K = 0, \qquad h_k = \sigma_k - \sigma_{k+1} > 0 \tag{5}$$

When the teacher and fake models are evaluated at the same state and noise level, their denoised-prediction difference satisfies

$$\widehat{\mathbf{x}}_{0,\phi} - \widehat{\mathbf{x}}_{0,\mathrm{T}} = \sigma\left(\mathbf{v}_{\mathrm{T}} - \mathbf{v}_{\phi}\right) \tag{6}$$

and their score difference is

$$s_T - s_\phi = -\frac{1-\sigma}{\sigma}\left(v_T - v_\phi\right), \quad \sigma > 0 \tag{7}$$

In practice, the teacher velocity v_T is produced by the frozen teacher and the fake velocity v_ϕ is produced by the online auxiliary model. The fake model is trained independently with a flow-matching objective on detached samples generated by the student.

3.2 On-Policy Teacher Anchoring

DMD provides distribution-level supervision, but does not explicitly constrain the local transport behavior of a preference-aligned teacher along the student's current few-step trajectory. We therefore anchor the student to the routed teacher at states visited by the student itself.

Let $\mathbf{x}_k$ be the source state of a student rollout trajectory at noise level σ_k. The student velocity at this state is

$$\mathbf{v}_k = \mathbf{v}_\theta(\mathbf{x}_k, \mathbf{c}) \tag{8}$$

The frozen teacher is queried at the same student-induced state to produce a local velocity target:

$$\mathbf{v}_{T,k} = \mathbf{v}_T(\mathbf{x}_k, \mathbf{c}) \tag{9}$$

The teacher output is detached from the computation graph. The on-policy teacher anchoring loss is

$$\mathcal{L}^{(k)}_{\text{anchor}} = \left\|\mathbf{v}_k - \mathrm{sg}\left[\mathbf{v}_{T,k}\right]\right\|_2^2 \tag{10}$$

Since $\mathbf{x}_k$ is generated by the current student sampler, the teacher is queried on student-induced states. This anchors the student's local velocity to the teacher at the states encountered during few-step sampling.

3.3 On-Policy Distribution Matching

In standard DMD, a clean student prediction is perturbed with Gaussian noise before querying the teacher and fake-score models:

$$\widetilde{\mathbf{x}}_\sigma = (1-\sigma)\mathbf{x}_0 + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{11}$$

At a queried state $\mathbf{x}$ and noise level σ, we define the detached DMD direction as the normalized teacher/fake velocity difference:

$$\mathbf{g}(\mathbf{x}, c) = \mathrm{sg}\left[\mathcal{N}_\sigma\left(v_T(\mathbf{x}, c) - v_\phi(\mathbf{x}, c)\right)\right] \tag{12}$$

Here, $\mathcal{N}_\sigma(\cdot)$ denotes the per-sample normalization used during training. We abbreviate $\mathbf{g}(\mathbf{x}, c)$ as $\mathbf{g}$. Its effect on the queried state can be represented by the following stop-gradient target loss:

$$\mathcal{L}_{\mathrm{DM}} = \frac{1}{2}\left\|\mathbf{x} - \mathrm{sg}[\mathbf{x} - \mathbf{g}]\right\|_2^2 \tag{13}$$

Since $\mathbf{g}$ is detached,

$$\nabla_{\mathbf{x}}\mathcal{L}_{\mathrm{DM}}(\mathbf{x}) = \mathbf{g} \tag{14}$$

Instead of evaluating the distribution-matching correction at an independently perturbed clean prediction, OPDMD evaluates it at a state induced by the student's current few-step rollout. For the k-th rollout state, the one-step successor state is

$$\mathbf{z}_k = \mathrm{sg}\left[\mathbf{x}_k\right] - h_k\mathbf{v}_k \tag{15}$$

Thus, $\mathbf{z}_k$ is the state reached by the student's rollout at noise level σ_{k+1}. The source state $\mathbf{x}_k$ is detached, while the dependence of $\mathbf{z}_k$ on $\mathbf{v}_k$ is retained for the gradient derivation. We query the

frozen teacher and the online fake-score model at $(\mathbf{z}_k, \mathbf{c})$. The velocity-space distribution-matching direction is

$$\mathbf{g}_k = \mathrm{sg}\left[\mathcal{N}_{\sigma_{k+1}}\left(v_T(\mathbf{z}_k, \sigma_{k+1}, c) - v_\phi(\mathbf{z}_k, \sigma_{k+1}, c)\right)\right] \tag{16}$$

To expose the gradient acting on the query state, we use the inner-product surrogate

$$\widetilde{\mathcal{L}}_{\mathrm{DM}}^{(k)} = \langle \mathbf{z}_k, \mathbf{g}_k \rangle \tag{17}$$

Because $\mathbf{g}_k$ is detached,

$$\nabla_{\mathbf{z}_k} \widetilde{\mathcal{L}}_{\mathrm{DM}}^{(k)} = \mathbf{g}_k$$

The inner-product surrogate and the stop-gradient target in Eq. 13 have different scalar values but induce the same gradient with respect to the queried state. The inner-product form is used only to expose the gradient path required for the derivation. It is not optimized as a separate loss in our implementation. Instead, its detached gradient is incorporated into a single MSE target together with the on-policy teacher anchor.

3.4 Transition-Aware MSE Loss

Because the rollout prefix $\mathbf{x}_k$ is detached, the query state in Eq. 15 depends on the student parameters only through $\mathbf{v}_k$. Therefore,

$$\frac{\partial \mathbf{z}_k}{\partial \mathbf{v}_k} = -h_k \mathbf{I} \tag{18}$$

Let $\lambda \geq 0$ denote the weight of the distribution-matching correction. Using the detached teacher target $\mathbf{v}_{T,k}$ and the detached DMD direction $\mathbf{g}_k$, the gradient of the separate anchor and DMD surrogate is

$$\nabla_{\mathbf{v}_k}\left(\mathcal{L}_{\mathrm{anchor}}^{(k)} + \lambda \widetilde{\mathcal{L}}_{\mathrm{DM}}^{(k)}\right) = 2\left(\mathbf{v}_k - \mathbf{v}_{T,k}\right) - \lambda h_k \mathbf{g}_k \tag{19}$$

We implement this gradient using a single detached target velocity:

$$\widetilde{\mathbf{v}}_k = \mathrm{sg}\left[\mathbf{v}_{T,k} + \frac{\lambda h_k}{2}\mathbf{g}_k\right] \tag{20}$$

and optimize the student with

$$\mathcal{L}_{\mathrm{OPDMD}}^{(k)} = \|\mathbf{v}_k - \widetilde{\mathbf{v}}_k\|_2^2 \tag{21}$$

The gradient of Eq. 21 is exactly the gradient in Eq. 19. This equivalence holds under the same state construction, mean reduction, and stop-gradient operations. The factor $h_k/2$ is determined by the Jacobian of the student-induced rollout transition and automatically accounts for the span of each sampling step. Consequently, OPDMD combines source-state teacher anchoring with distribution correction at the corresponding one-step successor state in a single transition-aware MSE update instead of optimizing OPD and DMD as two separate objectives.

4 Experiments

4.1 Experimental Setup

We evaluate preference-preserving few-step distillation on SD-3.5-Medium with multiple GRPO-aligned expert teachers (Liu et al., 2025). Each expert is optimized for a different capability, including text rendering, compositional generation, and preference alignment. During training, a prompt-level router assigns each prompt to the corresponding frozen expert teacher, while a single student is jointly distilled from all experts. We consider two challenging sampling settings, with 4 and 8 network function evaluations (NFE). Following Flow-OPD (Fang et al., 2026), we evaluate general compositional generation with GenEval (Ghosh et al., 2023), text rendering with OCR accuracy (Chen et al., 2023), and human preference alignment with PickScore (Kirstain et al., 2023). To quantify distributional drift after few-step distillation, we additionally compute the Fréchet Distance (FD) between student and teacher samples in DINOv2 (Oquab et al., 2023) feature space, where a lower value indicates a smaller distributional discrepancy. Compositional text–image alignment is further evaluated on T2I-CompBench++ (Huang et al., 2025). More training and implementation details are provided in Appendix A.

Table 1: Comparison of preference-aligned capability preservation under 4/8-step generation.

Model	NFE	FD ↓	OCR Acc. ↑	GenEval ↑	PickScore ↑
SD-3.5-M	35	152.73	0.5799	0.6384	21.58
OPD	35	90.07	0.8975	0.8882	22.72
+GRPO-OCR	35	-	0.9254	0.6619	21.80
+GRPO-GenEval	35	-	0.6450	0.9426	21.64
+GRPO-PickScore	35	-	0.6787	0.4828	23.18
OPD only	8	140.88	0.8591	0.8768	22.58
DMD only	8	157.92	0.8170	0.7800	22.61
OPD → DMD	8	160.97	0.8608	0.7768	22.44
DMD → OPD	8	146.81	0.8588	0.8627	22.73
Ours (const. weight)	8	105.48	0.8631	0.8816	22.70
Ours	8	**97.90**	**0.8639**	**0.9037**	**22.76**
OPD only	4	483.32	0.5363	0.6588	21.01
DMD only	4	145.37	0.8005	0.7836	22.63
OPD → DMD	4	143.55	0.8413	0.8313	22.60
DMD → OPD	4	471.30	0.5794	0.6820	21.06
Ours (const. weight)	4	180.10	0.8358	0.8746	22.32
Ours	4	**139.97**	**0.8835**	**0.8849**	**22.68**

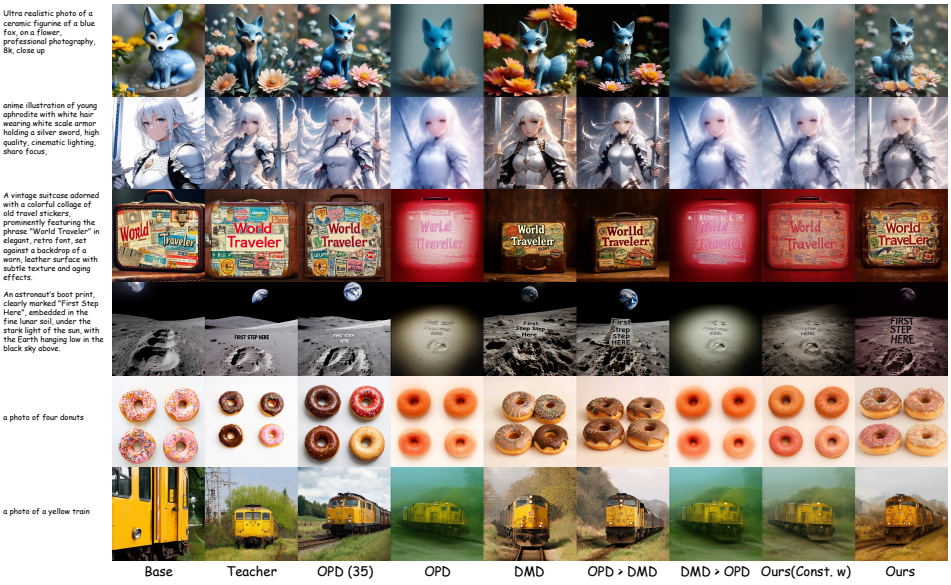

Figure 3: Qualitative comparison under 4-step generation.

4.2 Main Results

We compare OPDMD with OPD, DMD, and two sequential pipelines, OPD→DMD and DMD→OPD, under both 4-step and 8-step generation. As shown in Table 1, OPDMD achieves the best OCR accuracy, GenEval, and PickScore among all few-step methods. It also achieves the lowest FD, indicating a smaller distributional discrepancy from the teacher. The advantage becomes more pronounced under the more challenging 4-step setting. OPD suffers substantial degradation when trained on sparse trajectories, resulting in missing details and incomplete content, as illustrated in Fig. 3. DMD preserves sharper visual details but retains the teacher's aligned capabilities less effectively and exhibits noticeable changes in color and contrast. The two sequential variants alleviate some individual failure modes but do not consistently preserve both teacher-aligned capabilities and distributional fidelity. In contrast, OPDMD jointly preserves the teacher's aligned behavior and corrects the few-step student distribution within a single training stage, achieving better capability preservation and distributional fidelity. The qualitative comparisons in Figs. 3 and 4 further show that OPDMD better retains semantic details and the visual characteristics of the aligned teachers while avoiding the severe degradation in sparse-step OPD and sequential distillation.

We further evaluate compositional generation on T2I-CompBench++ (Huang et al., 2025), comparing OPDMD with DMD2 (Yin et al., 2024a), DMD2 without the GAN term, and Consistency

Table 2: Comparison on T2I-CompBench++ under 4-step and 8-step settings.

Model	NFE	Color	Shape	Texture	Complex	3D-Spatial	Numeracy	Non-Spatial
SD-3.5-M	40	0.7994	0.5669	0.7338	0.3800	0.3739	0.5927	0.3146
Flow-OPD	40	0.8298	0.6292	0.7446	0.3943	0.4565	0.6837	0.3163
DMD2 (w/o GAN)	8	0.8042	0.5955	0.7361	0.3808	**0.4536**	0.6766	0.3094
DMD2	8	0.8068	0.5949	0.7333	0.3804	0.4490	0.6643	0.3092
CD	8	0.7690	0.5674	0.6124	0.3243	0.2892	0.5337	0.3028
Ours	8	**0.8165**	**0.6126**	**0.7422**	**0.3893**	0.4465	**0.6781**	**0.3117**
DMD2 (w/o GAN)	4	0.8113	0.6017	**0.7406**	0.3849	0.4504	0.6709	0.3115
DMD2	4	0.8019	0.5934	0.7330	0.3858	**0.4507**	0.6658	0.3116
CD	4	0.7917	0.5743	0.6756	0.3495	0.3429	0.5797	0.3054
Ours	4	**0.8140**	**0.6072**	0.7321	**0.3931**	0.4388	**0.6749**	**0.3146**

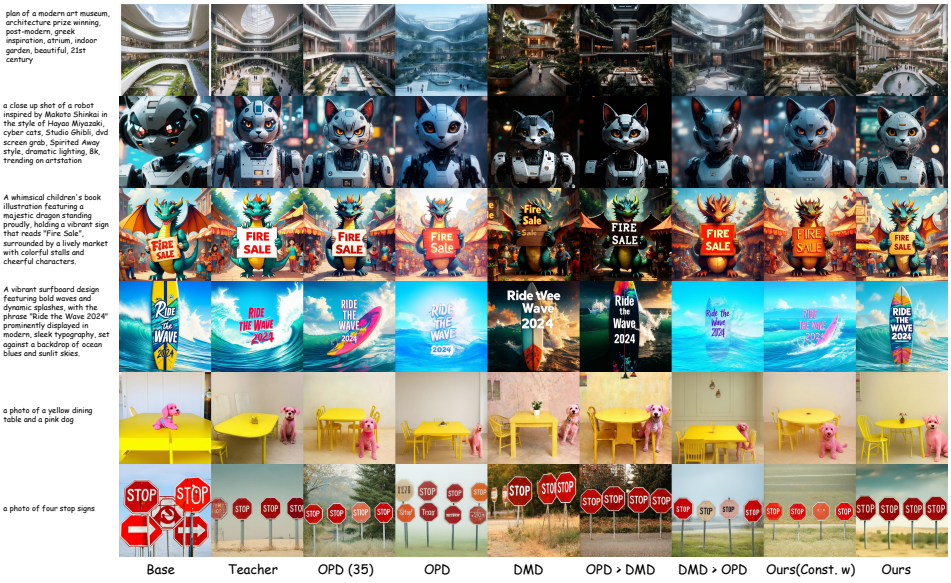

Figure 4: Qualitative comparison under 8-step generation.

Distillation (Song et al., 2023) (CD). Since the original real-image data are unavailable, teacher-generated images are used wherever image inputs are required during training. As shown in Table 2, OPDMD achieves the best result among the evaluated few-step methods on six of the seven dimensions at 8 steps and five of the seven dimensions at 4 steps. These results indicate that our method preserves strong compositional text-image alignment even with few sampling steps.

4.3 Analysis

One-Stage vs. Two-Stage Distillation Table 1, Figs. 3 and 4 show that neither sequential pipeline consistently balances capability preservation and few-step generation quality. OPD→DMD generally maintains image sharpness, but the subsequent DMD stage introduces noticeable changes in global appearance, including darker tones and increased contrast. Reversing the order leads to a different failure pattern: DMD→OPD remains relatively competitive at 8 steps but deteriorates substantially at 4 steps, where FD increases to 471.30 and its OCR accuracy and GenEval drop to 0.5794 and 0.6820, respectively. The corresponding samples also exhibit severe blurring and missing semantic details.

These observations indicate the mismatch between the two objectives. OPD constrains local teacher-consistent transport along the student trajectory, but sparse trajectory matching cannot compensate for sparse-step discretization error. DMD instead provides distribution-level correction without continuously constraining the local transport of the aligned teacher. When the two objectives are applied sequentially, the second stage can therefore partially overwrite properties established by the first. Our method avoids this stage-wise interference by applying both supervision signals throughout the same student rollout.

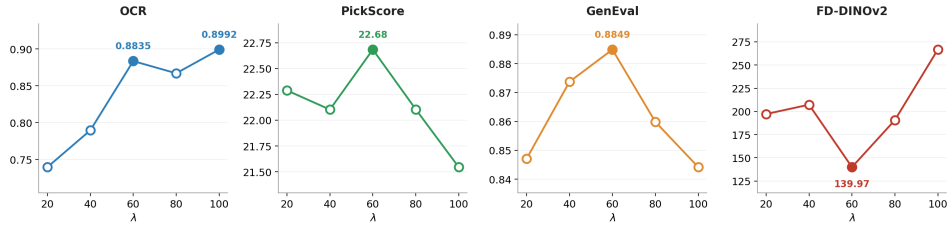

Figure 5: Effect of λ on 4-step OPDMD.

Transition-Aware vs. Constant Correction Weighting We evaluate the effect of transition-aware weighting by comparing OPDMD with a variant that applies a constant coefficient to the distribution-matching correction, denoted as **Ours (const. weight)** in Table 1. Both variants use the same teacher anchor, successor-state correction, and MSE formulation. They differ only in the correction coefficient: OPDMD uses $\lambda h_k/2$ at step k, whereas the baseline replaces h_k with the mean transition span $\bar{h} = K^{-1}\sum_{k=0}^{K-1} h_k$. The baseline therefore preserves the average correction coefficient while removing its dependence on the individual sampling transition. This distinction matters because the correction is evaluated at the successor state but applied to the source-state velocity. The mapping between these two quantities introduces the transition span h_k, which determines how strongly the successor-state correction contributes to the velocity update.

As shown in Table 1, transition-aware weighting improves all evaluated metrics under both 4-step and 8-step generation, with larger gains in the 4-step setting. At 4 steps, it reduces FD from 180.10 to 139.97 and improves OCR accuracy, GenEval, and PickScore from 0.8358, 0.8746, and 22.32 to 0.8835, 0.8849, and 22.68, respectively. These results suggest that accounting for the transition span helps balance local teacher anchoring and distribution correction across the rollout, particularly under sparse sampling.

Effect of λ. The coefficient λ controls the relative contribution of distribution matching in the transition-aware target. When $\lambda = 0$, the loss reduces to pure on-policy teacher anchoring; increasing λ strengthens the distribution-level correction, but overly large values can weaken the influence of the local teacher anchor. Fig. 5 shows the sensitivity to λ under 4-step generation. Among the evaluated settings, $\lambda = 60$ provides the most balanced performance, achieving the highest GenEval and PickScore together with the lowest FD. Although a larger value can improve the OCR accuracy, it degrades the overall performance and distributional fidelity. This trend highlights the complementary roles of the two supervision signals: insufficient distribution correction leaves the student vulnerable to sparse-step discretization error, whereas excessive correction can move the student away from the local behavior of the aligned teacher.

5 Conclusion

In this work, we introduced On-Policy Distribution Matching Distillation (OPDMD), a single-stage framework for distilling preference-aligned diffusion models into few-step generators. By jointly applying teacher anchoring and distribution matching along student-generated trajectories, OPDMD improves capability transfer when local teacher supervision on sparse trajectories is insufficient and mitigates the distributional drift that can arise during DMD-based few-step distillation. Its transition-aware objective combines the two supervision signals, scaling the distribution-matching correction by the sampling transition span to coordinate capability preservation and distribution correction throughout training. Experiments with GRPO-aligned SD-3.5-Medium teachers show that OPDMD improves capability preservation and distributional fidelity at both 4 and 8 steps. Ablations highlight the benefits of transition-aware correction under sparse sampling and the importance of balancing the two supervision signals. Together, these results support joint teacher anchoring and distribution matching for capability-preserving few-step generation. The current study focuses on text-to-image generation and relies on pretrained, preference-aligned teachers. Future work will explore jointly training the teacher and student within a single stage and extending the framework to video generation and other generative domains.

DISCLOSURE OF AI USE

We used OpenAI Codex to assist with manuscript editing, improve clarity and organization, and refine descriptions of the method and experimental results. We also used it to provide feedback on the consistency of mathematical notation, algorithm pseudocode, and experimental reporting, and to assist in interpreting experimental comparisons. The authors reviewed the AI-assisted revisions, checked technical descriptions against the method, and verified numerical statements against the experimental results. The authors take full responsibility for the final content of this paper, including all text, technical claims, and interpretations developed with AI assistance.

ETHICS STATEMENT

This work studies few-step distillation of preference-aligned text-to-image models. Reducing generation costs may benefit practical applications, but may also lower the barrier to producing misleading or harmful synthetic content. Our method aims to preserve the aligned capabilities of teacher models; it does not establish broader safety guarantees and may inherit biases or undesirable behaviors from the teachers and their training data. The reported improvements on capability and preference benchmarks should therefore not be interpreted as evidence of fairness or safety. Deployment requires application-specific safeguards and evaluation of potential misuse and bias.

REPRODUCIBILITY STATEMENT

We provide the formulation and derivation of OPDMD in Section 3 and summarize its training procedure in Algorithm 1. Section 4 describes the experimental setup, training settings, and evaluation metrics. The implementation code is included in the supplementary materials to facilitate reproduction of our results.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

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.

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.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 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 preprint arXiv:2605.05204*, 2026a.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, Changsheng Lu, Zhen Li, Mengmeng Wang, et al. Distribution matching distillation meets reinforcement learning. In *European Conference on Computer Vision*, pp. 281–299. Springer, 2026b.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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.

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

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 preprint arXiv:2604.13016*, 2026b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 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, 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728. IEEE, 2025.

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 preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685, 2022.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

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

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.

APPENDIX

A TRAINING DETAILS

Our experimental setup largely follows Flow-OPD (Fang et al., 2026). We use 16-step training rollouts for both the OPD baseline evaluated at 35 NFEs in Table 1 and the OPD stage of the OPD→DMD baselines. For OPDMD, the final-step correction is evaluated at the zero-noise successor state using the normalized teacher/fake velocity difference. This computation does not divide by the noise level, and the normalization denominator is clamped from below for numerical stability. For the OPD-only baselines evaluated at 4 and 8 NFEs, we use 4-step and 8-step training rollouts, respectively. Each training stage is performed for 2,000 steps at a resolution of 512×512. We also distill classifier-free guidance into the student, using a guidance scale of 4.5 during rollouts. The learning rates are 5×10^{-5} for the student and 1×10^{-4} for the online fake-score model. We perform five fake-model updates per student update. Both models use LoRA (Hu et al., 2021) with rank $r = 32$ and scaling parameter $\alpha = 64$. All experiments run on a machine with eight NVIDIA A100 80GB GPUs. For T2I-CompBench++ (Huang et al., 2025), we directly evaluate the publicly released Flow-OPD checkpoints and also use them for subsequent DMD and consistency distillation (CD). Following Flow-OPD (Fang et al., 2026), we report PickScore values scaled by 26, except in Fig. 1, where unscaled values are shown for comparison with the other metrics.

B TRAINING ALGORITHM

Algorithm 1 summarizes OPDMD training. Each iteration alternates between flow-matching updates to the online fake model on detached student samples and a student update using the transition-aware MSE objective. For the student update, we select a source state from a few-step rollout and detach the rollout prefix. The teacher provides a local velocity anchor at this state, while the normalized teacher–fake velocity difference at the one-step successor state provides the distribution-matching correction in Eq. 16. The detached target combines these signals with the factor $\lambda h_k/2$, following Eqs. 20 and 21. For each prompt, both teacher queries use the same frozen routed expert. In the algorithm, G_θ denotes the student's complete K-step sampler, and $\mathcal{N}_\sigma$ is the per-sample normalization. We suppress minibatch indices for readability and use mean reduction for the squared-error losses. Only θ and ϕ are updated; the teachers remain frozen, and the student update does not backpropagate through its target or rollout trajectory.

C MORE RESULTS

C.1 TEACHER SOLVER ABLATION

We compare Euler, DPM (Lu et al., 2022), Ralston (Karras et al., 2022), and Heun (von Platen et al., 2022) for constructing teacher velocity targets under the same 4-step OPDMD setting. As shown in Table 3, Euler achieves the lowest FD and the highest OCR accuracy and PickScore, while DPM yields a slightly higher GenEval score (0.8862 vs. 0.8849). The qualitative results in Fig. 6 show broadly similar outputs, with differences in text rendering, semantic details, and visual style. Overall, the higher-order alternatives offer no consistent improvement across metrics despite requiring additional teacher evaluations during training. We therefore use Euler in the main experiments for its balance of performance and computational cost.

C.2 DMD-DISTILLED TEACHERS FOLLOWED BY OPD

We further investigate whether the limitations of the DMD→OPD pipeline can be alleviated by applying DMD to the expert teachers before multi-teacher distillation. Specifically, we first distill each GRPO-aligned expert teacher into a task-specific few-step model using DMD. The resulting few-step GenEval, OCR, and PickScore experts are then used as routed teachers to train a single student with OPD. As shown in Table 4, the DMD-distilled experts retain relatively strong performance on their respective target tasks. However, combining their capabilities through a subsequent OPD stage leads to a clear degradation. Under the 8-step setting, the resulting OPD student obtains an OCR

Algorithm 1 Transition-Aware On-Policy Distribution Matching Distillation

Require: Student velocity model v_θ; online fake velocity model v_ϕ; frozen expert velocity models $\{v_{T_r}\}$; noise grid $1 = \sigma_0 > \cdots > \sigma_K = 0$; routed prompt dataset $\mathcal{D}$; optimizers $\mathcal{O}_\theta$ and $\mathcal{O}_\phi$; coefficient λ; per-sample normalization $\mathcal{N}_\sigma$; number of fake-model updates M

while not converged **do**
  Sample a minibatch of routed prompts $\{(\mathbf{c}_i, r_i)\}_{i=1}^{B} \sim \mathcal{D}$
  Select $v_T \leftarrow v_{T_r}$ for each prompt $\mathbf{c}$ ▷ Use the same frozen expert for both teacher queries
  // Update the online fake velocity model
  for $m = 1, \ldots, M$ **do**
    Sample $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Generate a detached student sample $\widehat{\mathbf{x}}_0 \leftarrow \mathrm{sg}[G_\theta(\boldsymbol{\eta}, \mathbf{c})]$
    Sample $\tau \sim \mathcal{U}(0,1)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{y} \leftarrow (1-\tau)\widehat{\mathbf{x}}_0 + \tau\boldsymbol{\epsilon}$
    $\mathbf{u} \leftarrow \boldsymbol{\epsilon} - \widehat{\mathbf{x}}_0$
    $\mathcal{L}_{\text{fake}} \leftarrow \|v_\phi(\mathbf{y}, \tau, \mathbf{c}) - \mathbf{u}\|_2^2$
    $\phi \leftarrow \mathcal{O}_\phi(\phi, \nabla_\phi \mathcal{L}_{\text{fake}})$ ▷ Update the online fake velocity model
  end for
  // Construct an on-policy student training state
  Sample $k \sim \mathrm{Uniform}\{0, \ldots, K-1\}$
  Sample $\mathbf{x}_{\sigma_0} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  // Roll out the student sampler to the selected node
  for $q = 0, \ldots, k-1$ **do**
    $\mathbf{x}_{\sigma_q} \leftarrow \mathrm{sg}[\mathbf{x}_{\sigma_q}]$
    $\mathbf{v}_q \leftarrow v_\theta(\mathbf{x}_{\sigma_q}, \sigma_q, \mathbf{c})$
    $h_q \leftarrow \sigma_q - \sigma_{q+1}$
    $\mathbf{x}_{\sigma_{q+1}} \leftarrow \mathrm{sg}\big[\mathbf{x}_{\sigma_q} - h_q \mathbf{v}_q\big]$ ▷ Perform one transition
  end for
  $\mathbf{x}_k \leftarrow \mathrm{sg}[\mathbf{x}_{\sigma_k}]$ ▷ Selected on-policy source state
  $\mathbf{v}_k \leftarrow v_\theta(\mathbf{x}_k, \sigma_k, \mathbf{c})$ ▷ Student velocity at the source state
  $h_k \leftarrow \sigma_k - \sigma_{k+1}$
  $\mathbf{z}_k \leftarrow \mathrm{sg}[\mathbf{x}_k] - h_k \mathbf{v}_k$ ▷ One-step successor state induced by the student
  $\mathbf{v}_{T,k} \leftarrow \mathrm{sg}[v_T(\mathbf{x}_k, \sigma_k, \mathbf{c})]$ ▷ Teacher velocity at the source state
  $\overline{\mathbf{z}}_k \leftarrow \mathrm{sg}[\mathbf{z}_k]$
  $\mathbf{v}_{T,k}^{\text{real}} \leftarrow \mathrm{sg}[v_T(\overline{\mathbf{z}}_k, \sigma_{k+1}, \mathbf{c})]$
  $\mathbf{v}_{\phi,k}^{\text{fake}} \leftarrow \mathrm{sg}[v_\phi(\overline{\mathbf{z}}_k, \sigma_{k+1}, \mathbf{c})]$
  $\mathbf{g}_k \leftarrow \mathrm{sg}\Big[\mathcal{N}_{\sigma_{k+1}}\Big(\mathbf{v}_{T,k}^{\text{real}} - \mathbf{v}_{\phi,k}^{\text{fake}}\Big)\Big]$ ▷ Normalized successor-state correction
  $\widetilde{\mathbf{v}}_k \leftarrow \mathrm{sg}\big[\mathbf{v}_{T,k} + \frac{\lambda h_k}{2}\mathbf{g}_k\big]$ ▷ Transition-aware target velocity
  $\mathcal{L}_{\text{OPDMD}}^{(k)} \leftarrow \|\mathbf{v}_k - \widetilde{\mathbf{v}}_k\|_2^2$
  $\theta \leftarrow \mathcal{O}_\theta\Big(\theta, \nabla_\theta \mathcal{L}_{\text{OPDMD}}^{(k)}\Big)$ ▷ Update the student model
end while

accuracy of 0.7497, a GenEval score of 0.7830, and a PickScore of 21.74, all lower than the corresponding task-specific DMD teachers. The degradation becomes more pronounced at 4 steps, where these scores decrease to 0.6454, 0.7479, and 20.82, respectively, while FD increases from 186.00 to 277.16. The qualitative comparisons in Figs. 7 and 8 further show that DMD introduces visual distribution shifts in the expert teachers, which the subsequent OPD stage transfers to the student, resulting in noticeable visual artifacts.

These results indicate that compressing the expert teachers before OPD does not resolve the mismatch between distribution matching and trajectory anchoring. DMD first modifies both the distributions and local behaviors of the task-specific teachers, after which sparse-trajectory OPD cannot reliably preserve and combine their aligned capabilities. In contrast, OPDMD jointly applies teacher anchoring and distribution correction to the original aligned teachers throughout training, avoiding the error accumulation caused by sequential distillation.

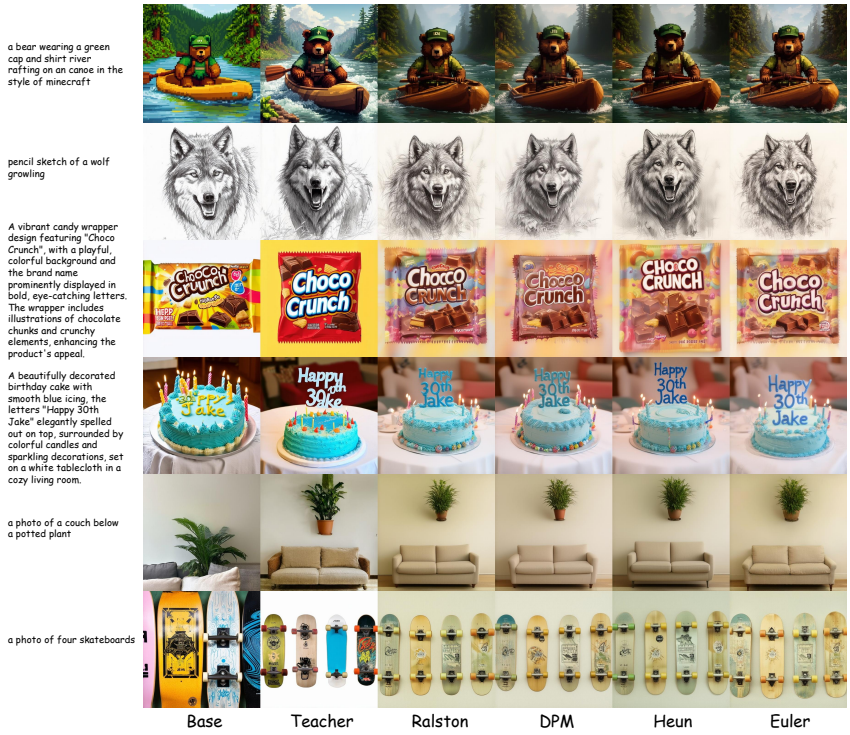

Figure 6: Qualitative comparison of different teacher solvers under the 4-step setting. Ralston, DPM, Heun, and Euler produce broadly comparable results, although minor differences appear in text rendering, semantic details, and visual style.

Table 3: Comparison of different teacher samplers for OPDMD under 4-step generation.

Teacher Sampler	NFE	FD ↓	OCR Acc. ↑	GenEval ↑	PickScore ↑
Euler	4	**139.97**	**0.8835**	0.8849	**22.68**
Ralston	4	140.50	0.8544	0.8778	22.40
DPM	4	140.55	0.8332	**0.8862**	22.46
Heun	4	164.00	0.8617	0.8743	22.32

Table 4: DMD-distilled expert teachers followed by multi-teacher OPD distillation. Each GRPO-aligned expert is first distilled into a task-specific 4-step or 8-step teacher using DMD. The three distilled experts are then used as routed teachers to train a single OPD student. Although the DMD teachers retain strong performance on their respective tasks, the subsequent OPD stage degrades aligned capabilities and increases the distributional discrepancy, especially under the 4-step setting.

Method	NFE	FD ↓	OCR Acc. ↑	GenEval ↑	PickScore ↑
SD-3.5-M	35	152.73	0.5799	0.6384	21.58
+GRPO-GenEval	35	-	0.6450	0.9426	21.64
+GRPO-OCR	35	-	0.9254	0.6619	21.80
+GRPO-PickScore	35	-	0.6787	0.4828	23.18
+GRPO-GenEval+DMD	8	246.77	0.4150	0.8849	21.40
+GRPO-OCR+DMD	8	215.47	0.7859	0.5621	21.56
+GRPO-PickScore+DMD	8	172.49	0.5630	0.3561	22.85
OPD	8	186.00	0.7497	0.7830	21.74
+GRPO-GenEval+DMD	4	239.74	0.4100	0.9121	21.52
+GRPO-OCR+DMD	4	212.70	0.8119	0.4412	21.74
+GRPO-PickScore+DMD	4	180.34	0.5241	0.3074	23.03
OPD	4	277.16	0.6454	0.7479	20.82

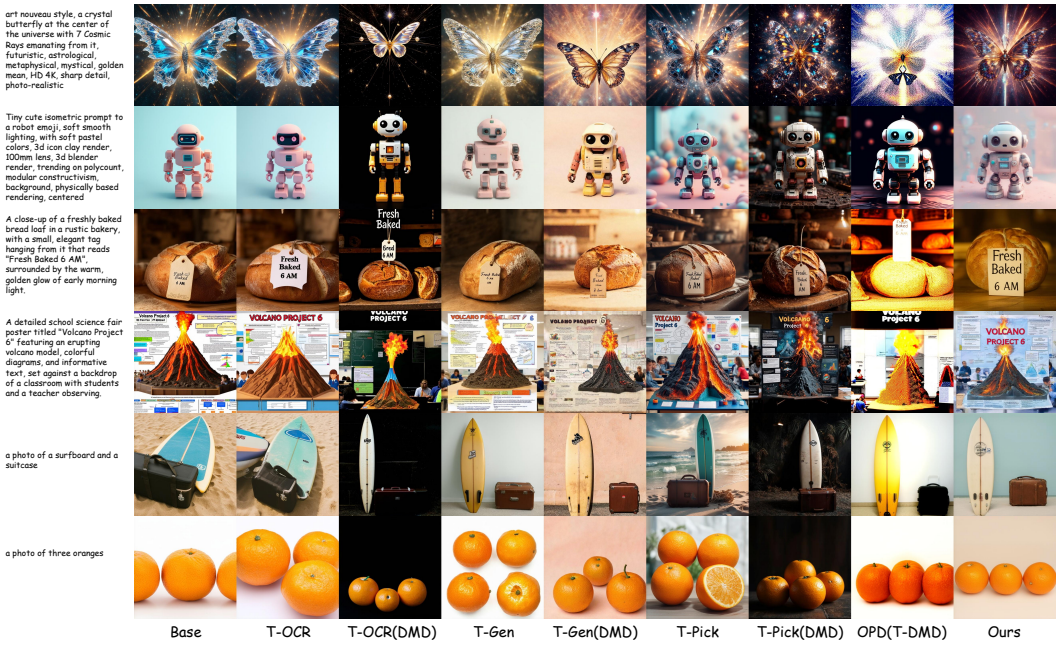

Figure 7: Qualitative comparison of DMD-distilled teachers followed by OPD under the 4-step setting. T-OCR, T-Gen, and T-Pick denote the original task-specific teachers, while T-OCR(DMD), T-Gen(DMD), and T-Pick(DMD) denote their DMD-distilled counterparts. OPD(T-DMD) is the student trained by OPD using these distilled teachers. Sequential distillation introduces noticeable changes in color and contrast and further degrades semantic and visual details.

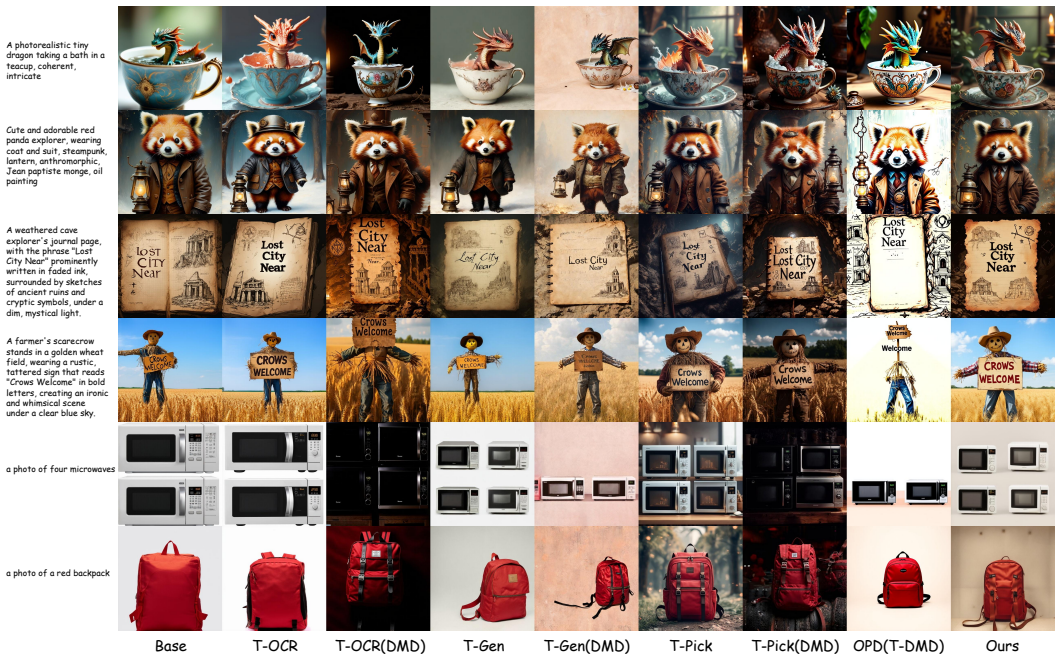

Figure 8: Qualitative comparison of DMD-distilled teachers followed by OPD under the 8-step setting. Although increasing the sampling steps improves the results of OPD(T-DMD), the sequential pipeline exhibits visual-distribution shifts and weaker preservation of teacher-aligned capabilities.