GFD-OPD: GUIDANCE-FOLDED ON-POLICY DISTILLATION OF DIFFUSION MODELS ACROSS SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has demonstrated two important capabilities in language models: compressing large teachers into smaller students and merging expert models into a single model. Existing diffusion OPD, however, mostly focus on the latter, with teachers and students sharing the same backbone and scale. We investigate large-to-small diffusion opd from large teachers to a small student and find that the standard recipe fails. To find the underlying cause, we propose **Fixed-State KL**, an effective and fair way to measure the distribution gap between student and teacher during OPD training for diffusion models. We are the first to clarify why large-to-small OPD is challenging for diffusion models: a smaller student struggles to perfectly match the distribution of a larger teacher, while classifier-free guidance can accumulate and amplify the distributional discrepancies between the student's conditional and unconditional branches and those of the teacher. To solve this problem, we propose **GFD-OPD**, a simple yet effective method that reduces the student–teacher gap while avoiding the error amplification of the CFG composition. Across numerous experiments, GFD outperforms previous baselines in both training efficiency and final performance, achieving state-of-the-art results on all benchmarks.

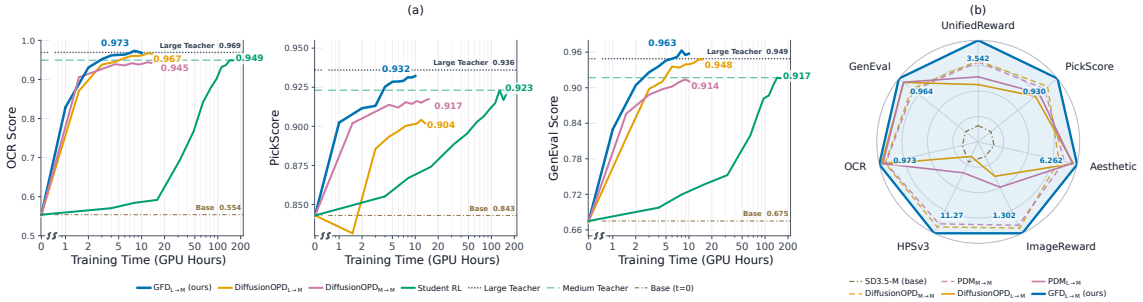

Figure 1: GFD-OPD distills SD3.5-Large into SD3.5-Medium. (a) Score versus training compute (GPU hours) on OCR, PickScore, and GenEval: GFD-OPD overtakes all baselines within 10 GPU-hours. (b) GFD-OPD ranks first on all seven evaluation dimensions, covering both task rewards and image quality.

1 INTRODUCTION

On-policy distillation (OPD) has recently emerged as a powerful post-training recipe for large language models. Like reinforcement learning (RL), OPD trains the student on trajectories sampled from its own policy and therefore optimizes the student's own distribution; unlike RL, which has to work with sparse and delayed rewards, OPD lets a teacher provide dense supervision at every step of the student-generated trajectory. This fine-grained feedback largely removes the credit-assignment difficulty of RL, so OPD enjoys a more stable optimization and a much better sample efficiency at a comparable final quality.

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.

In language models OPD is primarily applied in two paradigms. The first is *strong-to-weak* distillation, where the capability of a large teacher is transferred to a small student that is cheap to deploy (Lu & Lab, 2025). The second is *capability merging*, where several specialist models derived from the same base model are merged into a single student without conflicts (Zeng et al., 2026). Diffusion models have begun to adopt OPD as a post-training method as well (Li et al., 2026b; Fang et al., 2026), but the existing diffusion work mostly focus on the second paradigm. The first paradigm, distilling a large diffusion model into a small one, remains unexplored. It is also important: a compact model that inherits capabilities of a larger teacher can achieve near-teacher quality at inference while retaining a low cost close to that of the smaller model.

Our goal is to investigate effective approaches for distilling the capabilities of large diffusion models into smaller ones. Although DiffusionOPD performs effectively when merging models with the same architecture, we find directly extending it to large-to-small distillation leads to severe degradation. As shown in Fig. 4, generated images exhibit artifacts, such as corrupted regions, loss of fine-grained textures and so on.

To understand the source of this degradation, we propose **Fixed-State KL** for measuring the distributional discrepancy between the student and teacher. Specifically, we compute KL divergence on states sampled from normal teacher trajectories and noise-corrupted trajectories constructed from real images, providing a fair and well-behaved evaluation distribution for all methods.

Under this evaluation metric, we find that large-to-small DiffusionOPD produces a substantially larger distributional discrepancy between the student and teacher than its same-architecture counterpart. Furthermore, we quantitatively characterize models' final sampling result through spectral energy analysis of the final latent representations and high-pass texture analysis of the decoded images, which quantify the deviation from the teacher at both coarse-grained (e.g., global tone) and fine-grained (e.g., local texture) levels.

Through extensive analysis, we indicate that a smaller-capacity student consistently struggles to perfectly match the distribution of a larger teacher. Besides, classifier-free guidance (CFG), used during diffusion sampling, can accumulate and amplify the distributional discrepancies between the student's conditional and unconditional branches and those of the teacher, causing a larger deviation in the final generation trajectory.

Motivated by these observations, we propose a simple yet effective method, **Guidance Folding Distillation (GFD)**. GFD reduces the distributional discrepancy between the student and teacher while avoiding the error amplification introduced by CFG composition, enabling effective large-to-small OPD for diffusion models.

Our contributions are as follows.

- We propose **Fixed-State KL**, an effective and fair way to measure the distribution gap between student and teacher during OPD training for diffusion models.
- Through multifaceted analysis, we are the first to clarify **why large-to-small OPD is challenging for diffusion models**. To address these challenges, we propose **GFD**, a simple and effective method that reduces the student–teacher gap while avoiding the error amplification of the CFG composition.
- Through multiple experiments, combined with quantitative and qualitative analysis, we show the effectiveness of GFD. GFD removes the artifacts in cross-scale DiffusionOPD and outperforms previous baselines in both training efficiency and final performance.

2 METHOD

2.1 PRELIMINARY: ON-POLICY DISTILLATION IN LANGUAGE MODELS AND DIFFUSION MODELS

On-policy distillation (OPD) trains the student on trajectories sampled from its own policy, while the teacher provides dense supervision on every visited state. In language models, OPD minimizes the reverse KL

between the student policy $\pi_{\boldsymbol{\theta}}$ and the frozen teacher $\pi^\star$ along student-generated sequences:

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{LLM}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}}\Big[\sum_{t=1}^{|\boldsymbol{y}|} \mathrm{KL}\big(\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{y}_{<t}) \,\|\, \pi^\star(\cdot \mid \boldsymbol{y}_{<t})\big)\Big]. \tag{1}$$

DiffusionOPD transfers the same idea to diffusion and flow-matching samplers. At each denoising step j, the student and teacher induce Gaussian transition kernels $p_S(x_{t_{j+1}} \mid x_{t_j})$ and $p_T(x_{t_{j+1}} \mid x_{t_j})$ whose covariance $\sigma_j^2 \boldsymbol{I}$ is fixed by the shared sampler, so the per-step reverse KL has the closed form $\|\mu_S(x_{t_j};\boldsymbol{\theta}) - \mu_T(x_{t_j})\|_2^2/(2\sigma_j^2)$. Since the transition mean is affine in the velocity field with sampler-fixed coefficients, and sampling deploys the CFG-guided velocity, DiffusionOPD matches the CFG-guided output to the teacher's:

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{diff}}(\boldsymbol{\theta}) = \mathbb{E}_{x_{0:N}\sim p_{S,\boldsymbol{\theta}}}\left[\sum_{j=0}^{N-1} \omega_j \big\| v_S^g(x_{t_j}, t_j, \boldsymbol{c}) - v_T^g(x_{t_j}, t_j, \boldsymbol{c})\big\|_2^2\right], \tag{2}$$

where $v_\star^g = v_\star^u + w\,(v_\star^c - v_\star^u)$, with $\star \in \{S, T\}$ indexing the student and the teacher, is the guided velocity composed from the conditional and unconditional branches $v_\star^c, v_\star^u$ at guidance scale w, and $\omega_j = \eta_j^2/(2\sigma_j^2)$ is fixed by the sampler, with η_j the coefficient of the velocity in the transition mean and σ_j^2 the per-step variance; $\eta_j = \big(1+\sigma_{t_j}^2(1-t_j)/(2t_j)\big)\Delta t_j$ and $\sigma_j^2 = \sigma_{t_j}^2 \Delta t_j$. In both SDE and ODE settings, DiffusionOPD is thus a pathwise, closed-form velocity-matching problem along the student's rollout.

2.2 Fixed-state KL Evaluation

We first directly apply the standard DiffusionOPD objective to the large-to-small OPD setting. However, after training, from Fig. 4 we can see that the resulting images exhibit obvious visual artifacts: they contain corrupted textures and are generally darker. To further quantify the degradation in generation quality, we conduct a spectral energy analysis on the final latent representations and a high-pass texture analysis on the decoded images. As shown in Fig. 6, DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$ deviates substantially from the teacher across all evaluated aspects.

To investigate the underlying cause, we need to measure the distributional discrepancy between the student and teacher. A natural choice is to use the objective in OPD training: the reverse KL divergence between the student and teacher distributions, computed along the trajectory rolled out by the student. However, although this formulation is well suited as a training objective, we argue that it is not an appropriate metric to evaluate the distributional discrepancy between the student and the teacher after training.

There are two main reasons. First, different training methods generally produce different student policies. Computing the KL divergence on each student's own trajectories therefore evaluates the models on different sets of states, confounding the effect of distributional mismatch with differences in state visitation. It makes the resulting metrics difficult to compare in a fair and controlled manner.

Second, the student model may deviate from the normal generation trajectory during training and enter abnormal states. As pointed out by Xin et al. (2026), the teacher is no longer guaranteed to provide reliable predictions in such states, since the queried states may lie far outside the region on which the teacher's behavior is meaningful. Consequently, the KL divergence computed on these abnormal states may no longer faithfully reflect the discrepancy between a well-behaved student and the teacher.

Motivated by these considerations, we propose **Fixed-State KL**, which is performed on a fixed and normal set of states shared by all student models. Specifically, we use states from normal teacher trajectories, together with noise-corrupted trajectories constructed from real images, as common evaluation inputs for computing the KL divergence between the student and teacher. By decoupling the evaluation states from the student's own rollout distribution, this protocol enables a more controlled and comparable assessment of how closely different students match the teacher distribution.

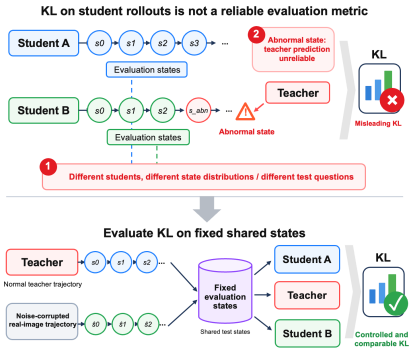

Figure 2: Illustration of the KL evaluation.

Table 1: KL divergence between each student and its corresponding teacher. $\mathrm{P_{stu}}$ and $\mathrm{P_{tea}}$ denote sampling trajectories generated by the student and teacher models, respectively, while $\mathrm{P_{real}}$ denotes trajectories constructed by adding noise to real images.

	KL ($\mathrm{P_{tea}}$)	KL ($\mathrm{P_{real}}$)	KL ($\mathrm{P_{stu}}$)	branch KL ($\mathrm{P_{tea}}$)	
Student	guided	guided	guided	cond	uncond
$\mathrm{DiffusionOPD_{M\to M}}$	0.041	0.022	0.039	0.017	0.023
$\mathrm{DiffusionOPD_{L\to M}}$	0.068	0.046	0.057	0.037	0.039
$\mathrm{PDM_{L\to M}}$	0.063	0.046	0.052	0.023	0.022
$\mathrm{split\text{-}KL_{L\to M}}$	0.074	0.049	0.065	**0.021**	**0.020**
$\mathrm{GFD_{L\to M}}$ $(w=1)$	**0.050**	**0.030**	**0.043**	–	–
$\mathrm{GFD_{L\to M}}$ $(w=2)$	0.053	0.032	0.053	–	–

Furthermore, as shown in Tab. 1, the KL divergence evaluated on student rollout trajectories exhibits a markedly different trend from that computed on teacher rollout trajectories and noise-corrupted trajectories constructed from real images. The former fails to reflect the true trend of student–teacher distributional discrepancy. It may provide a misleading evaluation signal.

2.3 From Cross-scale DiffusionOPD to GFD

Based on the KL evaluation above, we can figure out why the performance of DiffusionOPD in $L \to M$ setting degrade severely. As shown in Table 1, the guided-output KL, conditional KL, and unconditional KL of $\mathrm{DiffusionOPD_{L\to M}}$ are all higher than those of $\mathrm{DiffusionOPD_{M\to M}}$. "guided output" refers to the CFG-combined output used during sampling. Accordingly, the final evaluation should focus on the distributional KL divergence between the student's guided output and the teacher's guided output.

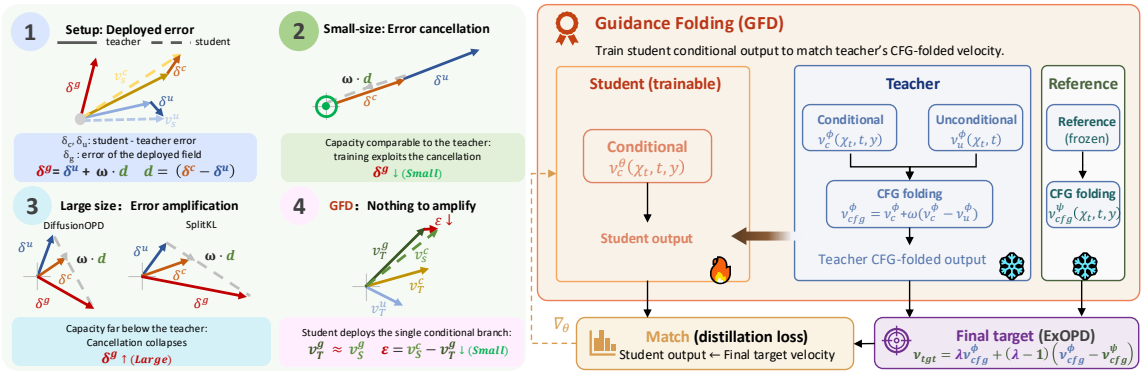

Figure 3: **GFD-OPD overview. Left** (motivation): the deployed error of a CFG-composed student decomposes as $\delta_g = \delta_u + \omega\, d$ — same-size distillation survives through error cancellation, large-to-small does not, and guidance folding removes the composition so that nothing is left for ω to amplify. **Right** (method): the process of GFD-OPD.

Attempting branch-wise supervision Recall that DiffusionOPD objective directly matches the student's guided prediction v_S^g to the teacher's guided prediction v_T^g. It lacks separately constraining of the conditional and unconditional branches. Such an objective may be sufficient when the student has comparable model capacity to the teacher, as in the $M \to M$ setting. However, in a large-to-small setting, the smaller student has limited approximation capacity, which may make the conditional and unconditional branches more prone to

drifting away from their teacher counterparts. Such branch-wise drift may in turn lead to a larger discrepancy in the final guided output.

This observation motivates us to explicitly supervise the conditional and unconditional branches separately, thereby preventing them from drifting. A natural choice is to adopt the following training objective:

$$\mathcal{L} = \|v_S^c - v_T^c\|_2^2 + \|v_S^u - v_T^u\|_2^2, \tag{3}$$

PDM (Li et al., 2026a) also propose a similar objective as

$$\mathcal{L} = \|v_S^c - v_T^c\|_2^2 + \alpha \|v_S^g - v_T^g\|_2^2, \tag{4}$$

However, their motivation is fundamentally different from ours: we use this objective to prevent the conditional and unconditional branches from drifting in the large-to-small setting but theirs are not.

After training with separate supervision on the two branches, from Table 1, we observe that the KL divergences of both the conditional and unconditional branches are indeed reduced. However, the KL divergence of the final guided output does not decrease accordingly. Consistent with this observation, the generated images in Fig. 4 still exhibit noticeable quality degradation. As shown in Fig. 6, the corresponding quantitative metrics also remain substantially deviated from those of the teacher.



Figure 4: Shared cross-architecture artifacts under a SD3.5-L teacher. Rows: the SD3.5-M base, three prior OPD recipes (*DiffusionOPD*, *PDM*, *splitKL*), GFD and large teacher.

Why the branch error induce but the guided field does not? To investigate the cause of the increased guided-output KL, we decompose the error of the guided prediction as

$$\delta^g = \delta^u + wd, \tag{5}$$

where $\delta^g = v_S^g - v_T^g, \delta^u = v_S^u - v_T^u, \delta^c = v_S^c - v_T^c$, and $d = \delta^c - \delta^u$. Table 6 reveals that, when trained to directly match the guided output, DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$ is able to exploit the cancellation between the two error components, δ^u and wd, resulting in a relatively small final guided error δ^g. In contrast, for DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$, the limited approximation capacity of the smaller student weakens such error cancellation. For PDM and SplitKL, separately supervising the conditional and unconditional branches largely

removes this cancellation effect, causing the errors to be amplified rather than canceled after guidance composition. We provide a more detailed analysis in Section B.1.

This observation further suggests that, in the large-to-small setting, discrepancies in the conditional and unconditional branches can be systematically amplified by the guidance operation, ultimately leading to a larger shift in the final generative distribution.

Guidance Folding for Large-to-Small Distillation This motivates a different design: directly absorb the teacher's guided policy into the student's conditional branch, eliminating the need to recover the target distribution through the composition of two imperfectly matched branches. We propose **Guidance Folding Distillation (GFD)**. Rather than jointly optimizing the student's conditional and unconditional distributions, it converts the original two-branch composition problem into a single-branch velocity matching problem.

Formally, GFD minimize the discrepancy between the student's conditional distribution and the teacher's guided distribution:

$$\min_{\theta} \mathbb{E}_{x_t \sim d^S} \left[D_{\mathrm{KL}} \left(\pi_S^c(\cdot \mid x_t, c) \,\|\, \pi_T^g(\cdot \mid x_t, c) \right) \right]. \tag{6}$$

Applying the per-step derivation of Eq. 2, the GFD objective can be written explicitly as

$$\mathcal{L}_{\mathrm{GFD}}(\boldsymbol{\theta}) = \mathbb{E}_{x_{0:N} \sim p_{S,\boldsymbol{\theta}}} \left[\sum_{j=0}^{N-1} \omega_j \left\| v_S^c(x_{t_j}, t_j, \boldsymbol{c}; \boldsymbol{\theta}) - v_T^g(x_{t_j}, t_j; w) \right\|_2^2 \right]. \tag{7}$$

Table 1 shows that GFD achieves substantially lower KL divergence than the other methods. As illustrated in Fig. 4 and Fig. 6, GFD also restores normal generation quality, and the generated images are closer to those of the teacher in both coarse-grained and fine-grained metrics.

But a small, systematic gap remains. Fig. 6 shows that GFD ($w = 1$) has slightly weaker spectral energy than the teacher's guided output. And Appendix B.2 shows that the student learns the direction of the teacher's field but shrinks its magnitude toward its own prior. At inference this can be patched with a slightly larger guidance scale, with $w = 2$ the student distribution moves closer to the teacher's. GFD does not constrain the unconditional branch during training. In the low-noise regime, the unconditional branch co-evolves with the conditional branch and exhibits similar shifts, resulting in the CFG only containing a small increment and avoiding excessive guidance. Although the trajectory-level KL increases slightly at $w = 2$, we attribute this to the limited sensitivity of this metric to differences at such a small scale.

2.4 Reward Extrapolation for Large-to-Small Distillation

Yang et al. (2026) show that OPD in language models is a KL-regularized RL objective in which the teacher enters through a dense reward, and that scaling this reward by a factor $\lambda > 1$ moves the distillation target away from the reference model and beyond the teacher. We adopt this idea to further improve the student: by extrapolating the target in the direction away from the student's prior, the trained field is brought closer to the teacher without relying on an inference-time guidance increment.

In language models, the ExOPD objective is equivalent to standard OPD with the teacher replaced by the target $\pi_{\mathrm{tgt}} \propto (\pi^\star)^\lambda (\pi_{\mathrm{ref}})^{1-\lambda}$, where π_{ref} is a frozen reference policy. We apply the same replacement to the per-step transition kernels of diffusion OPD: $p_{\mathrm{tgt}} \propto p_T^\lambda \, p_R^{1-\lambda}$, where p_R is the kernel induced by the reference velocity v_R. Since both kernels are Gaussians with the same sampler-fixed variance, p_{tgt} remains a valid Gaussian kernel with that variance for every λ, and only its mean moves, corresponding to the extrapolated velocity (Appendix A.1):

$$v_{\mathrm{tgt}} = \lambda \, v_T + (1 - \lambda) \, v_R = v_T + (\lambda - 1)(v_T - v_R). \tag{8}$$

Table 2: SD3.5 merged multi-objective OPD results.

Model	Task rewards			Image quality			
	GenEval	OCR	Pick Score	Aes-thetic	Image Reward	Unified Reward	HPSv3
SD3.5-M (base)	0.675	0.554	0.843	5.373	0.907	3.201	9.020
SD3.5-L (base)	0.720	0.716	0.854	5.527	1.025	3.289	9.853
Per-task Large teachers							
GenEval-Teacher	0.949	0.801	0.834	5.463	1.144	3.416	9.812
OCR-Teacher	0.700	0.969	0.850	5.405	1.057	3.298	8.982
PickScore-Teacher	0.781	0.688	0.936	6.245	1.300	3.535	11.29
Per-task Medium teachers							
GenEval-Teacher	0.917	0.558	0.845	5.336	1.074	3.332	9.024
OCR-Teacher	0.599	0.948	0.824	5.129	0.749	3.071	6.345
PickScore-Teacher	0.696	0.605	0.924	6.188	1.283	3.460	11.55
DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$	0.908	0.944	0.917	6.067	1.275	3.461	11.07
PDM$_{\mathrm{M}\to\mathrm{M}}$	0.902	0.940	0.913	6.124	1.261	3.455	10.97
DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$	0.943	0.962	0.901	6.222	1.009	3.365	8.844
PDM$_{\mathrm{L}\to\mathrm{M}}$	0.944	0.955	0.906	6.211	1.065	3.396	9.360
split-KL$_{\mathrm{L}\to\mathrm{M}}$	0.939	0.959	0.912	6.211	1.103	3.417	10.08
GFD$_{\mathrm{L}\to\mathrm{M}}$(Ours)	**0.963**	**0.973**	**0.930**	**6.262**	**1.302**	**3.542**	**11.27**

Table 3: SD3.5 single-objective OPD results.

Model	Task rewards			Image quality
	GenEval	OCR	Pick Score	HPSv3
SD3.5-Medium (base)	0.675	0.554	0.843	9.020
Teachers (SD3.5-L)	0.949	0.969	0.936	–
DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$				
GenEval	0.951	–	–	6.844
OCR	–	0.967	–	6.097
PickScore	–	–	0.904	8.377
GFD $_{\mathrm{L}\to\mathrm{M}}$*(Ours)*				
GenEval	**0.963**	–	–	9.128
OCR	–	**0.973**	–	8.755
PickScore	–	–	**0.932**	**11.24**

Table 4: FLUX.2 merged OPD results.

Model	Task rewards			Image quality
	LongText	GenEval	Pick Score	HPSv3
FLUX.2-4B (base)	0.624	0.769	0.823	9.214
Teacher (FLUX.2-9B)	0.891	0.820	0.839	9.629
PDM$_{\mathrm{L}\to\mathrm{M}}$	0.635	0.815	0.829	8.904
DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$	0.635	0.830	0.832	9.180
GFD$_{\mathrm{L}\to\mathrm{M}}$(Ours)	**0.692**	**0.880**	**0.839**	**10.20**
Teacher (FLUX.2-32B)	0.979	0.893	0.862	11.29
PDM$_{\mathrm{L}\to\mathrm{M}}$	0.763	0.887	0.849	10.17
DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$	0.828	0.885	0.856	**10.69**
GFD$_{\mathrm{L}\to\mathrm{M}}$(Ours)	**0.857**	**0.908**	**0.859**	10.47

3 Experiments

3.1 Experimental Setup

Tasks and evaluation. We follow the three-reward suite of Flow-GRPO (Liu et al., 2026) and reuse its official prompt splits: GenEval (Ghosh et al., 2023) for compositional alignment, OCR accuracy for visual text rendering, and PickScore (Kirstain et al., 2023) optimized on Pick-a-Pic prompts and evaluated on DrawBench (Saharia et al., 2022). Task rewards are the primary metrics; following Flow-GRPO, we also report Aesthetic, ImageReward (Xu et al., 2023), UnifiedReward, and HPSv3 on DrawBench as image-quality probes. All methods are evaluated at 1024×1024 with the same prompts, seeds, and sampler, each at its own deployment guidance scale ($w = 4.5$ for CFG-composition students, $w = 2$ for GFD-OPD; §2).

Baselines. (a) The pretrained SD3.5-Medium and SD3.5-Large base models (Esser et al., 2024); (b) direct Flow-GRPO on both, which also serve as our teachers; (c) DiffusionOPD, PDM, and the split-KL variant of §2, trained with the same teachers, student initialization, and prompts as ours.

Implementation details. All OPD methods share the same on-policy setup: following Fast Flow-GRPO, rollouts use a 10-step reverse-SDE sampler with a 2-step training window. Training uses full-parameter fine-tuning at 1024×1024 with learning rate 5×10^{-6} and batch size 48 on 8 H100 GPUs; each single-task GFD-OPD run finishes in about 10 GPU-hours, versus 15 for DiffusionOPD and 160 for direct Flow-GRPO (Fig. 1). GFD-OPD composes the folded target at $w = 2$ with extrapolation $\lambda = 1.25$.

3.2 Main Results

Large-to-small distillation on SD3.5 and FLUX.2. Table 3 reports the single-objective SD3.5 setting (Large→Medium). GFD-OPD outperforms DiffusionOPD on all three rewards, surpasses the teacher on GenEval (0.962 vs. 0.949) and OCR (0.973 vs. 0.969), and is the only student whose HPSv3 stays at or above the untrained base—DiffusionOPD falls clearly below it, the fidelity loss behind the artifacts of §2.

In the merged multi-objective setting (Table 2), the baselines trade one side for the other: with the medium teacher they preserve image quality but are capped by its rewards; with the large teacher their rewards rise but quality drops (haze, desaturation, soft glyphs; Fig. 4). GFD-OPD attains the best task rewards (0.964 GenEval, 0.973 OCR, 0.930 PickScore) and the best image quality at the same time; per §3.3, folding alone roughly matches the teacher and the extrapolated target accounts for the rest. The same picture holds on FLUX.2 (Table 4; 9B and 32B teachers into 4B, OCR replaced by our long-text dataset): GFD-OPD leads on all task rewards, again exceeds the teacher on GenEval, and widens its LongText margin at 32B (qualitative examples in Fig. 8, appendix).

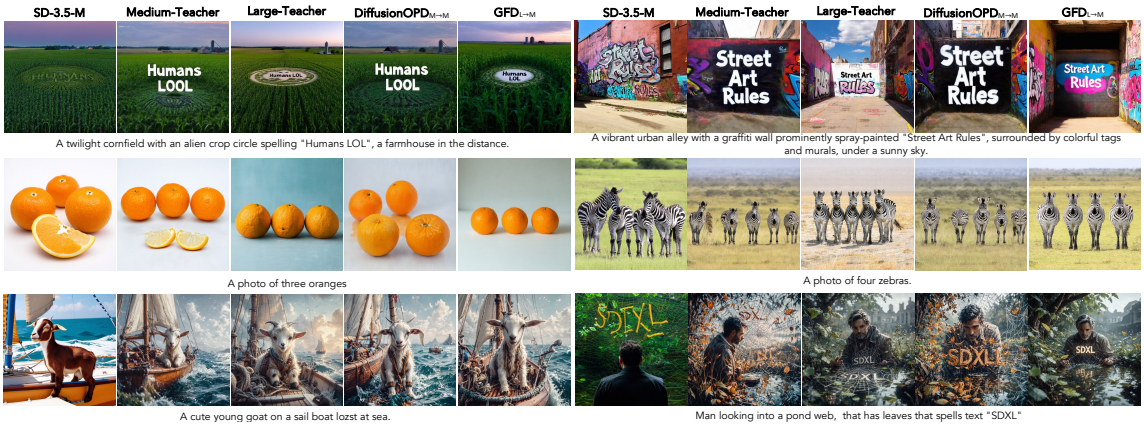

Figure 5: Qualitative comparison at 1024×1024. Columns, left to right: SD3.5-M base; the medium and large teachers; DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$; and GFD$_{\mathrm{L}\to\mathrm{M}}$.

Qualitative comparison. Fig. 5 presents a qualitative comparison. Since the large-teacher composition students already suffer from the artifacts shown in Fig. 4, we compare with DiffusionOPD distilled from the medium teacher, together with the base model and the two RL teachers. These medium-teacher students are free of visible artifacts, but their prompt adherence and fine-grained rendering are weaker: the crop-circle text in "Humans LOL" loses glyph structure, and the orange scene misses part of the required objects. Among all compared methods, GFD-OPD is the closest to the large teacher in both layout and fine detail. Besides, Fig. 8 shows qualitative examples for the FLUX.2 experiments of Table 4.

3.3 Ablations

Component ablation Table 5 adds the two components on top of standard OPD one at a time. Standard OPD improves the base on all three rewards but stays well below the teacher. Guidance folding contributes most of the gain, removing the artifacts of and bringing the student roughly to the teacher level; the extrapolated target adds a further consistent gain and pushes GenEval and OCR past the teacher. This supports the attribution: folding closes the gap to the teacher, and extrapolation provides the margin beyond it.

Table 5: Ablation: standard OPD + guidance folding + extrapolated-mean target (full GFD-OPD), on the SD3.5-Medium student at 1024×1024; HPSv3 probes image quality.

Method	GenEval	OCR	PickScore	HPSv3
SD3.5-Medium (base)	0.675	0.554	0.843	9.020
Standard OPD	0.884	0.908	0.880	9.175
+ Guidance folding	0.952	0.954	0.923	11.19
+ Extrapolated mean (GFD-OPD)	**0.963**	**0.973**	**0.930**	**11.27**

The impact of CFG scale In Fig. 6, we evaluate different models under varying CFG scales w, measuring the discrepancy between their final sampling results and those of their respective teacher models. DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$ approaches the origin at $w = 4.5$, indicating a close match to its teacher. In contrast, for DiffusionOPD, PDM, and Split-KL under the $\mathrm{L}\to\mathrm{M}$ setting, no choice of w brings the results close to the origin. Their outputs remain substantially different from those of the teacher. GFD achieves the closest match to the teacher at $w = 2$.

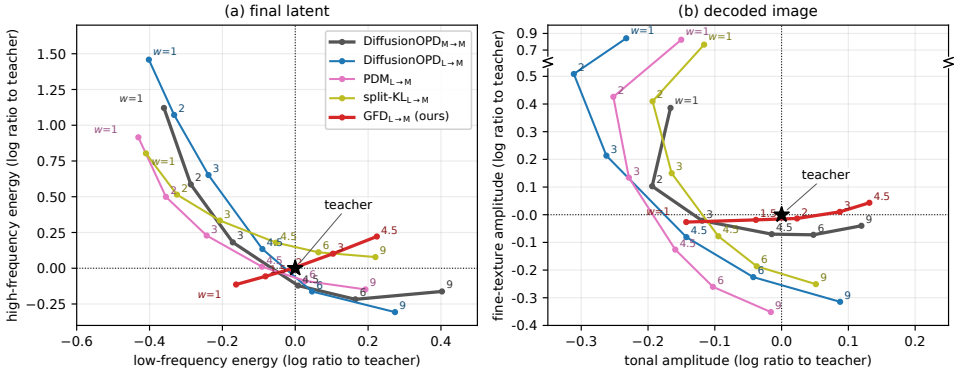

Figure 6: Statistics of the student's final output as the guidance scale w varies. Each axis is the log ratio of a statistic of the student's samples to the teacher's, so the teacher sits at the origin ($\star$). **(a)** Final latent: low-frequency energy (global tone) vs. high-frequency energy (fine texture). **(b)** Decoded image: tonal amplitude (RMS contrast $\times$ saturation) vs. fine-texture amplitude (3px high-pass MAE). The definition of all metrics are as detailed in Appendix B.3.

4 Related Work

On-Policy Distillation. OPD (Agarwal et al., 2024; Gu et al., 2024; Lu & Lab, 2025) post-trains language models by aligning the student with the teacher on student-generated trajectories, and is used both to compress large teachers into small students and to merge domain specialists into a single model. Yang et al. (2026) further interpret OPD as dense KL-constrained RL, we adopt this view in §2.4.

On-Policy Distillation for Diffusion Models. A recent line of work ports OPD to diffusion and flow-matching models (Ma et al., 2026; Fang et al., 2026; Wei et al., 2026). DiffusionOPD (Li et al., 2026b) derives the closed-form per-step objective and merges task-specific teachers into a unified student; PDM (Li et al., 2026a) supervises the conditional and unconditional CFG branches separately. These methods target same-family merging, while the large-to-small regime we study remains largely unexplored.

5 Conclusion

We studied on-policy distillation of diffusion models in the large-to-small regime and showed that the existing recipe does not transfer. To diagnose this we propose Fixed-State KL, which fairly makes methods comparable. The analysis attributes the failure to the CFG composition: a small student cannot fit the teacher's guided field exactly, the composition amplifies the error of its guidance increment by the guidance scale. GFD-OPD directly absorb the teacher's guided policy into the student's conditional branch. It reduces the student–teacher gap while avoiding the error amplification of the CFG composition. We hope this study can encourage further investigation into how the distributional discrepancy between teacher and student affect OPD of diffusion models.

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.

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.

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

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.

Bingnan Li, Haozhe Wang, Haozhong Xiong, Fangtai Wu, Jinpeng Yu, Yang Shi, Jiaming Liu, and Ruihua Huang. Rethinking classifier-free guidance in on-policy diffusion distillation. *arXiv preprint arXiv:2607.24731*, 2026a.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

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

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

Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the kl agreement trap in on-policy distillation. *arXiv preprint arXiv:2606.09471*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

A Derivation

A.1 Reward extrapolation and the extrapolated target

Reward extrapolation in language models. Yang et al. (2026) generalize the OPD objective (Eq. 1) by introducing a frozen reference policy π_{ref} and a reward weight λ:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\Big[\lambda \log \frac{\pi^\star(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}|\boldsymbol{x})}\Big] - \mathrm{KL}\big(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}) \,\|\, \pi_{\text{ref}}(\cdot|\boldsymbol{x})\big). \tag{9}$$

Standard OPD is the case $\lambda = 1$ (the reference cancels). Define $\pi_{\text{tgt}}(\boldsymbol{y}|\boldsymbol{x}) = (\pi^\star)^\lambda(\pi_{\text{ref}})^{1-\lambda}/Z(\boldsymbol{x})$ with $Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}}(\pi^\star)^\lambda(\pi_{\text{ref}})^{1-\lambda}$. Expanding the KL term,

$$\begin{aligned} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\big[\lambda \log \pi^\star - \lambda \log \pi_{\text{ref}} - \log \pi_{\boldsymbol{\theta}} + \log \pi_{\text{ref}}\big] \\ &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\big[\lambda \log \pi^\star + (1-\lambda)\log \pi_{\text{ref}} - \log \pi_{\boldsymbol{\theta}}\big] = -\,\mathrm{KL}\big(\pi_{\boldsymbol{\theta}} \,\|\, \pi_{\text{tgt}}\big) + \log Z(\boldsymbol{x}). \end{aligned} \tag{10}$$

Since $Z(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$, maximizing J is the same as minimizing the reverse KL to π_{tgt}; that is, ExOPD is standard OPD with the teacher replaced by π_{tgt}, where

$$\log \pi_{\text{tgt}} = \log \pi^\star + (\lambda - 1)\big(\log \pi^\star - \log \pi_{\text{ref}}\big) + \text{const}. \tag{11}$$

For $0 < \lambda < 1$ the target interpolates between reference and teacher; for $\lambda > 1$ it extrapolates along the reference→teacher direction.

Diffusion. At denoising step j, the teacher and reference kernels are $p_T = \mathcal{N}(\boldsymbol{\mu}_T, \sigma_j^2\boldsymbol{I})$ and $p_R = \mathcal{N}(\boldsymbol{\mu}_R, \sigma_j^2\boldsymbol{I})$ with the same sampler-fixed variance (§2.1). Applying the replacement above to the kernels gives $p_{\text{tgt}} \propto p_T^\lambda p_R^{1-\lambda}$.

Lemma 1. *Let $p_1 = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\boldsymbol{I})$ and $p_2 = \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2\boldsymbol{I})$. For every $\lambda \in \mathbb{R}$, the normalized product $p \propto p_1^\lambda p_2^{1-\lambda}$ equals $\mathcal{N}\big(\lambda\boldsymbol{\mu}_1 + (1-\lambda)\boldsymbol{\mu}_2,\ \sigma^2\boldsymbol{I}\big)$.*

Proof. $\log p(\boldsymbol{x}) \overset{c}{=} -\frac{\lambda}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_1\|_2^2 - \frac{1-\lambda}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_2\|_2^2$. The coefficient of $\|\boldsymbol{x}\|_2^2$ is $-\frac{\lambda+(1-\lambda)}{2\sigma^2} = -\frac{1}{2\sigma^2}$, independent of λ, so the covariance is $\sigma^2\boldsymbol{I}$; the linear term is $\frac{1}{\sigma^2}\langle \boldsymbol{x}, \lambda\boldsymbol{\mu}_1 + (1-\lambda)\boldsymbol{\mu}_2\rangle$, and completing the square gives the mean. □

By Lemma 1, $p_{\text{tgt}} = \mathcal{N}\big(\lambda\boldsymbol{\mu}_T + (1-\lambda)\boldsymbol{\mu}_R,\ \sigma_j^2\boldsymbol{I}\big)$: extrapolation changes the mean only, and p_{tgt} is a valid transition kernel for every λ, including $\lambda > 1$. Because the transition mean is affine in the velocity with sampler-fixed coefficients, $\boldsymbol{\mu} = \alpha_j\boldsymbol{x}_{t_j} + \beta_j v$, the target mean corresponds to $v_{\text{tgt}} = \lambda v_T + (1-\lambda)v_R$ (Eq. 8). The reverse KL between the student kernel and p_{tgt} is a KL between Gaussians with equal covariance, $\|\boldsymbol{\mu}_S - \boldsymbol{\mu}_{\text{tgt}}\|_2^2/(2\sigma_j^2)$, which in velocity space is the weighted matching loss of Eq. 2 with the target velocity v_{tgt} in place of v_T^g.

B Detailed Analysis

B.1 Analysis of error cancellation

The two error components. Given the guided output

$$v_g = v_u + w(v_c - v_u),$$

the guided prediction error can be written as

$$\delta^g = \delta^u + wd,$$

where

$$d = \delta_c - \delta_u.$$

Alternatively, d can be expressed as

$$d = \Delta_S - \Delta_T,$$

where

$$\Delta_S = v_S^c - v_S^u, \qquad \Delta_T = v_T^c - v_T^u.$$

These two error components have distinct meanings. Specifically, δ^u represents the prompt-independent denoising prediction error, while d captures the error in the prompt-induced increment.

Table 6: Decomposition $\boldsymbol{\delta}^g = \boldsymbol{\delta}^{\mathrm{u}} + w\boldsymbol{d}$ of the guided error at $w=4.5$. Cross term$= 2\langle \boldsymbol{\delta}^{\mathrm{u}}, 4.5\boldsymbol{d}\rangle / \|\boldsymbol{r}\|^2$.

Student	$\|\boldsymbol{\delta}^{\mathrm{u}}\|$	$\|\boldsymbol{d}\|$	$\|\boldsymbol{r}\|$	cross term
DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$ (reference)	0.092	0.032	**0.106**	$\mathbf{-1.56}$
DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$	0.090	0.043	0.177	-0.48
PDM$_{\mathrm{L}\to\mathrm{M}}$	**0.053**	0.041	0.178	-0.18
split-KL$_{\mathrm{L}\to\mathrm{M}}$	**0.051**	0.043	0.185	-0.18

Cancellation in $\mathrm{M}\to\mathrm{M}$ and its loss in $\mathrm{L}\to\mathrm{M}$. The guided objective only constrains the sum of the two components, $\delta^u + wd$. Therefore, any pair of errors satisfying $d = -\delta^u/w$ produces the same guided prediction error and lies in the null space of the objective. Consequently, minimizing this objective does not necessarily force both error components to vanish simultaneously. Instead, it allows the student to learn a negative correlation between d and δ^u, enabling one component to compensate for the other.

DiffusionOPD$_{\mathrm{M}\to\mathrm{M}}$ nearly reaches such a configuration, where the two error components δ^u and d substantially cancel each other. For DiffusionOPD$_{\mathrm{L}\to\mathrm{M}}$, besides a similar shrinkage effect, the error in d contains additional components that are orthogonal to δ^u, resulting in weaker error cancellation.

PDM and Split-KL explicitly supervise the two branches separately. Although this reduces the magnitude of the unconditional error component δ^u, it also makes the two error components nearly orthogonal to each other. As a result, the cancellation effect is largely removed, and the final guided error remains substantial.

B.2 Shrinkage of the trained field along the teacher's guided field

Tab. 7 quantifies the residual shrinkage of the guidance-folded student and how the inference-time guidance increment compensates it. At every step we project the student's deployed field onto the teacher's guided field, $g = \langle \boldsymbol{v}_S(w), \boldsymbol{v}_T^{\mathrm{cfg}}\rangle / \|\boldsymbol{v}_T^{\mathrm{cfg}}\|^2$, so that $g=1$ means the student carries the teacher's full amplitude along the teacher's direction. Results in the table is averaged over 500 prompts and over the steps of each window.

Table 7: Projection g of the student's deployed field onto the teacher's guided field.

Student	Probe	0–9	10–19	20–29	30–37	0–37 (total)
GFD ($w=1$)	$\mathrm{P}_{\mathrm{tea}}$	0.963	0.984	0.991	0.983	0.980
	$\mathrm{P}_{\mathrm{real}}$	0.943	0.971	0.987	0.984	0.968
GFD ($w=2$)	$\mathrm{P}_{\mathrm{tea}}$	1.030	0.991	0.993	0.984	**1.000**
	$\mathrm{P}_{\mathrm{real}}$	1.023	0.978	0.989	0.985	**0.997**

At $w = 1$ the trained field of GFD is short along $\boldsymbol{v}_T^{\mathrm{cfg}}$ in every window, most so at the high-noise steps 0–9. The student has thus learned the direction of the teacher's guidance but shrinks its magnitude toward

its own prior. Raising the guidance scale to $w = 2$ compensates the shrinkage, and the average over steps 0–37 moves from 0.97–0.98 to 1.00 on both probe sets. Besides, in the low-noise regime, the unconditional branch co-evolves with the conditional branch and exhibits similar shifts, leading to only a marginal increase of the CFG increment and avoiding excessive guidance.

B.3 Metrics of Fig. 6

All four quantities compare a student sample with the teacher sample generated from the same prompt and the same initial noise, over $N = 500$ prompts; each axis of Fig. 6 is a log ratio to the teacher, so the teacher sits at the origin.

Latent frequency energies (panel a). Let $z \in \mathbb{R}^{64\times64\times64}$ be the final latent in the model's packed token layout (64 channels on a 64×64 grid) and $\hat{z}_c = \mathrm{DCT}_2(z_c)$ the orthonormal 2-D DCT-II of channel c. With $u, v \in \{0, \frac{1}{63}, \dots, 1\}$ the normalized frequency indices and $\rho(u, v) = \sqrt{u^2 + v^2}/\sqrt{2} \in [0, 1]$ the radial frequency, the energy of a band B is

$$E_B(z) = \frac{1}{64^3} \sum_c \sum_{(u,v)\in B} \hat{z}_c(u, v)^2, \qquad B_{\text{low}} = \{\rho < 0.25\}, \quad B_{\text{high}} = \{\rho \geq 0.75\}. \tag{12}$$

The two coordinates of panel (a) are $\log\big(\frac{1}{N}\sum_n E_B(z_n^S)/E_B(z_n^T)\big)$ for $B = B_{\text{low}}$ (horizontal) and $B = B_{\text{high}}$ (vertical), where z_n^S and z_n^T are the student's and the teacher's final latents for prompt n.

Image tonal and fine-texture amplitude (panel b). The decoded 1024^2 image I is area-downsampled to 512^2 and converted to luminance $Y = 0.2126R + 0.7152G + 0.0722B$. Tonal amplitude combines the RMS contrast and the mean saturation,

$$c(I) = \frac{\mathrm{std}(Y)}{\mathrm{mean}(Y)}, \qquad s(I) = \operatorname*{mean}_p \frac{\max_k I_k(p) - \min_k I_k(p)}{\max_k I_k(p)}, \qquad a_{\text{tone}}(I) = \sqrt{c(I)\, s(I)}, \tag{13}$$

with k ranging over the RGB channels and p over pixels. Fine-texture amplitude is the mean magnitude of the 3×3 high-pass residual of the luminance,

$$a_{\text{tex}}(I) = \operatorname*{mean}_{p\in\Omega} \big|Y(p) - \mathrm{box}_{3\times3} Y(p)\big|, \tag{14}$$

where $\mathrm{box}_{3\times3}$ is the 3×3 mean filter with reflect padding and Ω excludes a 16-pixel border. The two coordinates of panel (b) are the mean log ratios $\frac{1}{N}\sum_n \log\big(a_{\text{tone}}(I_n^S)/a_{\text{tone}}(I_n^T)\big)$ (horizontal) and $\frac{1}{N}\sum_n \log\big(a_{\text{tex}}(I_n^S)/a_{\text{tex}}(I_n^T)\big)$ (vertical).

C More Ablations

C.1 GFD with unconditonal branch supervision

Supervising the unconditional branch (GFD+uncond). GFD only constrains the conditional branch and leaves the unconditional branch unconstrained. A natural variant, GFD+uncond, keeps the folded target for the conditional branch and additionally matches the student's unconditional branch to the teacher's:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{x_{0:N}\sim p_{S,\boldsymbol{\theta}}} \left[\sum_{j=0}^{N-1} \omega_j \Big(\|v_S^c - v_T^g\|_2^2 + \|v_S^u - v_T^u\|_2^2 \Big) \right]. \tag{15}$$

Same starting point. At $w = 1$ the two students deploy fields trained toward the same target and are indistinguishable: their KL divergence values are very similar, and their final sampling results sit at the same point in Fig. 7.

Wrong increment direction. What changes is the content of the CFG increment $v_S^c - v_S^u$. In GFD the unconditional branch co-evolves with the conditional branch and exhibits similar shifts, resulting in the CFG only containing a small increment and avoiding excessive guidance. Raising w mainly adds a native-scale texture correction. However, GFD+uncond matches the unconditional branch to v_T^u, removing the co-evolving. Its CFG increment is consequently doubled in magnitude. As demonstrated in Tab. 8 and Fig. 7, this causes a larger KL divergence and drives the final sampled results further away from the teacher distribution.

Table 8: KL divergence to the teacher for GFD and GFD+uncond, in the format of Tab. 1.

Student	KL (P_{tea})	KL (P_{real})
$GFD_{L\to M}$ ($w=1$)	0.050	0.030
$GFD_{L\to M}$ ($w=2$)	0.053	0.032
$GFD+uncond_{L\to M}$ ($w=1$)	0.048	0.029
$GFD+uncond_{L\to M}$ ($w=2$)	0.060	0.040

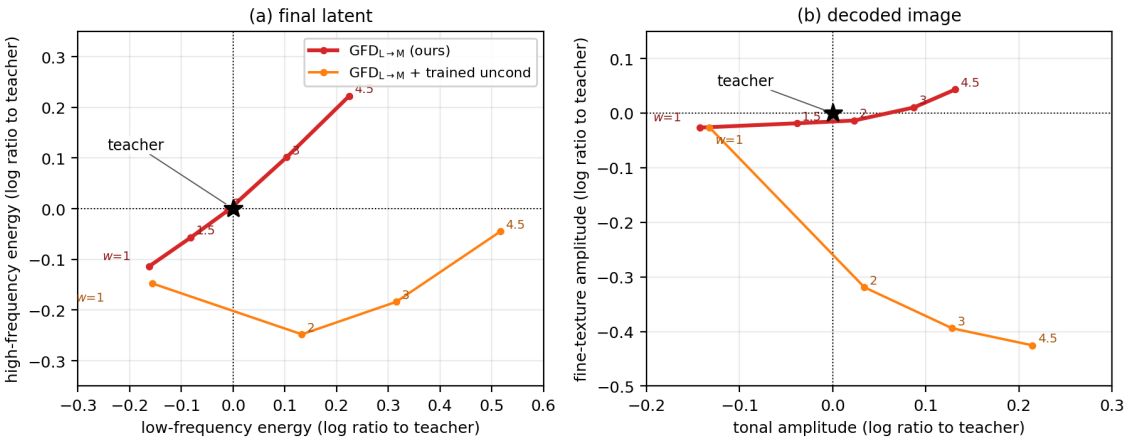

Figure 7: Statistics of the final output of GFD and GFD+uncond as the guidance scale w varies, in the coordinates of Fig. 6 (log ratio to the teacher, teacher at the origin).

D ADDITIONAL QUALITATIVE RESULTS ON FLUX.2

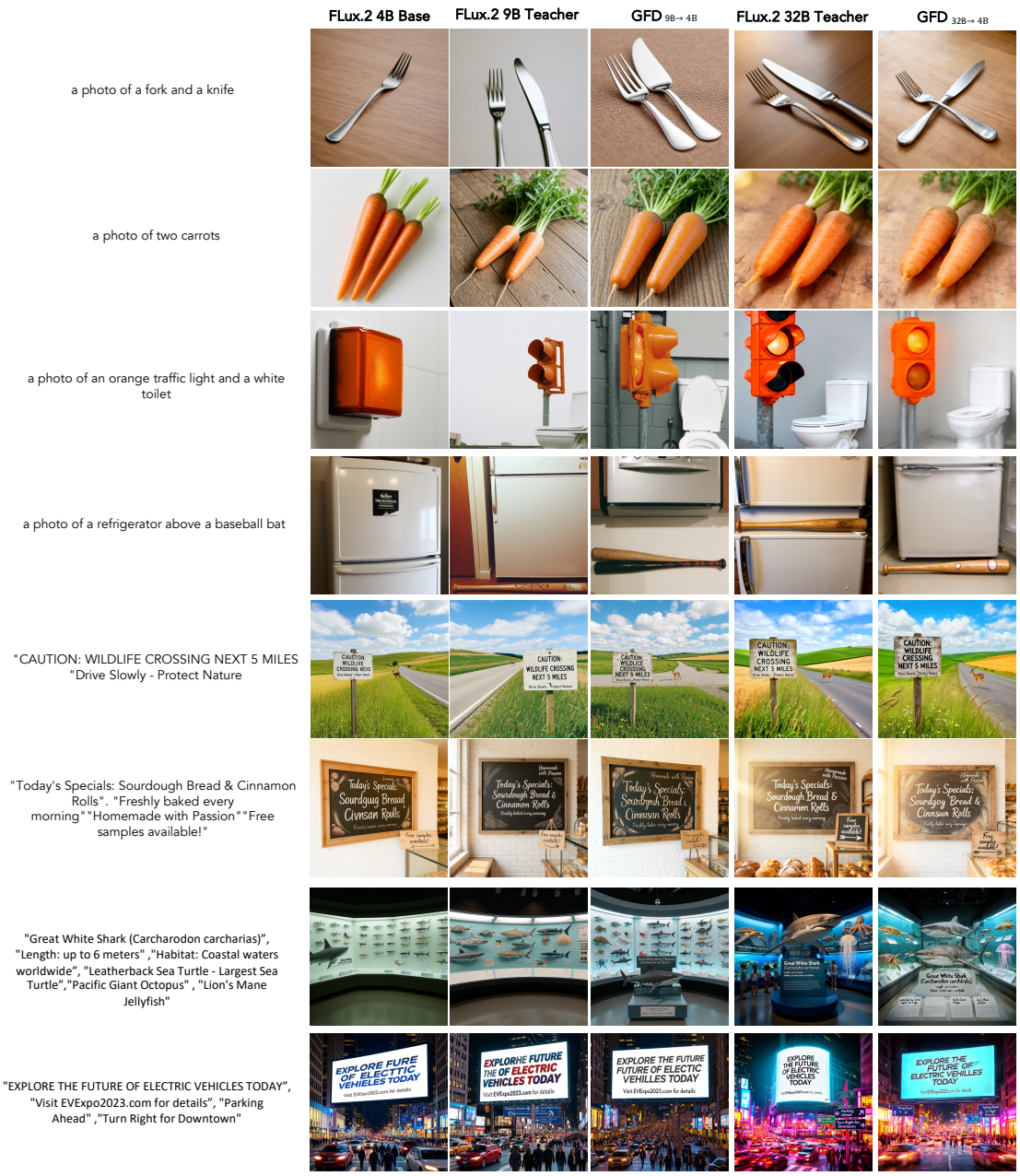

Figure 8: Qualitative comparison on FLUX.2 at 1024×1024, complementing Table 4. Columns, left to right: the FLUX.2-4B base model; the FLUX.2-9B teacher and the GFD-OPD student distilled from it into 4B; the FLUX.2-32B teacher and the GFD-OPD student distilled from it into 4B. The top four rows are GenEval prompts; the bottom four are LongText prompts probing long-form visual text rendering.