CONSISTENT DISTRIBUTION MATCHING FOR DATA-FREE DIFFUSION DISTILLATION

Anonymous authors
Paper under double-blind review

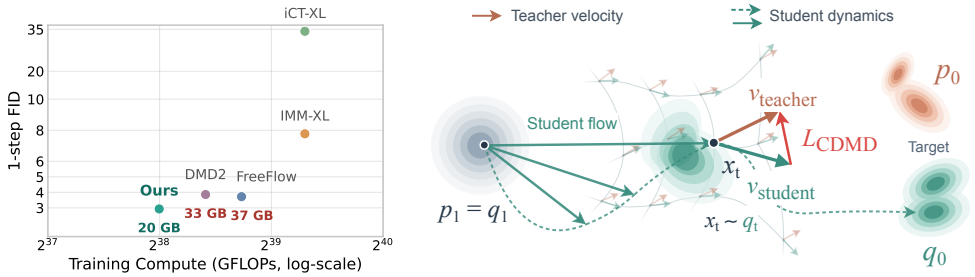


Figure 1: Left: compute versus FID-50K on ImageNet 256×256. The training memory below is measured with the same configuration on a single A100 GPU. Right: Consistent Distribution Matching overview. Our method aligns the teacher and student marginal probability paths without data or simulation.

ABSTRACT

Flow and diffusion models suffer from slow inference due to computationally expensive numerical integration. Distillation provides a promising way for a student model to learn from a teacher's dynamics, enabling one-step or few-step generation. However, existing methods often depend on curated distillation datasets, costly teacher rollouts, or auxiliary proxy networks, which complicate model training and scaling. In this work, we propose Consistent Distribution Matching, a simulation-free and data-free distillation method for accelerating diffusion and flow models while preserving strong generative capacity. Our key insight is to unify sample generation and score estimation with one student network. Thus, our framework uses only **two** models, a frozen teacher and a trainable student, and optimizes **one** objective. We prove that minimizing our objective indicates Wasserstein convergence of the student flow-map pushforwards to the teacher marginals. On ImageNet 256×256, our method attains an FID of **2.05** with a single function evaluation (1-NFE) and a 4-NFE FID of **1.37** within 40 epochs of training, surpassing the state-of-the-art distillation baselines without data.

1 INTRODUCTION

Diffusion and flow-based generative models (Albergo et al., 2023; Lipman et al., 2023; Liu et al., 2023; Dhariwal & Nichol, 2021; Song et al., 2021; Karras et al., 2022; Ho et al., 2020; Rombach et al., 2022; Tong et al., 2025) have become a leading framework for high-fidelity image and video synthesis. By learning a continuous transport from a simple prior distribution to a complex data distribution, these models provide stable training and strong sample quality across a wide spectrum of generative tasks (Fu et al., 2025; Yan et al., 2023; Wu et al., 2023b; Li et al., 2025; Zhang et al., 2025; Yadav et al., 2025; Yan et al., 2025; Xu* et al., 2025). Despite these advantages, their inference cost remains non-trivial. The primary bottleneck lies in the fact that generating a single sample requires numerically solving the differential equations governing the generative dynamics through an iterative denoising procedure, which often involves hundreds of neural function evaluations.

To address this challenge, diffusion distillation (Geng et al., 2024; Song et al., 2023; Song & Dhariwal, 2023; Lu & Song, 2024; Meng et al., 2023; Frans et al., 2024; Sabour et al., 2025; Fu et al., 2025) provides a natural route to accelerate the generation with a new model. Given a powerful teacher

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.



Figure 2: Curated 1024×1024 samples produced by our 4-step CDMD generator distilled from SANA1.6B LORA (Xie et al., 2024a). Please zoom in for details.

diffuser, the student is trained to reproduce the teacher's generative behavior while compressing the inference process in one or a few steps. Existing distillation methods have made substantial progress, yet their training pipelines often introduce additional complexity. Some depend on heavy backward simulation (Yin et al., 2024a; Gu et al., 2023) of a teacher diffusion model, auxiliary fake score networks (Yin et al., 2024a; Xie et al., 2024b; Zhou et al., 2024b), adversarial objectives (Zhou et al., 2025d; Kim et al., 2023; Sauer et al., 2024; 2025), or multi-stage training (Lee et al., 2024). Although these components can improve empirical performance, they incur substantial memory overhead, impose an egregiously high optimization burden, and ultimately hinder the scalability of the distillation procedure.

Beyond this procedural complexity, reliance on external distillation data (Yin et al., 2024b; Peng et al., 2025; Song et al., 2023; Lu & Song, 2024; Meng et al., 2023; Zhou et al., 2024b) introduces a fundamental source of mismatch between data and teacher probability path. Data-based methods construct intermediate states by noising data samples, but these states generally differ from those visited by the teacher during generation. As a result, the student is supervised on a surrogate state distribution rather than the teacher's actual probability path. This mismatch of the supervision states (Tong et al., 2025) becomes pronounced when fine-tuning tilts the teacher's distribution away from the original data (Xie et al., 2024b; Chen et al., 2025; Liu et al., 2025) or when the teacher adopts classifier-free guidance (Ho & Salimans, 2022) or autoguidance (Karras et al., 2024). In these settings, noising samples from the original data distribution provides progressively less faithful coverage of the states visited by the teacher. Moreover, when a teacher model is trained on proprietary data (Rombach et al., 2022; Wu et al., 2025; Seedream et al., 2025; Labs, 2025), constructing such a data-anchored supervision path may be infeasible. Direct supervision on teacher trajectories would address the mismatch, but obtaining such states requires backward simulation, thereby reintroducing the aforementioned computational burden. This trade-off motivates our central question: *can we distill a diffusion or flow teacher into a few-step student without external data, teacher rollouts, or auxiliary fake score models?*

We propose Consistent Distribution Matching Distillation (CDMD), a data-free and simulation-free framework for distilling diffusion and flow models using only a frozen teacher and a trainable student. Our framework circumvents the distribution mismatch issues. Our central idea is to unify sample generation and score estimation within a single flow map model, eliminating the need for a separately trained proxy network. The student predicts average velocities over arbitrary denoising time intervals, supporting generation in one or a few steps. We bootstrap intermediate training states from student predictions and analytic forward noising, avoiding real data and teacher rollouts while allowing the

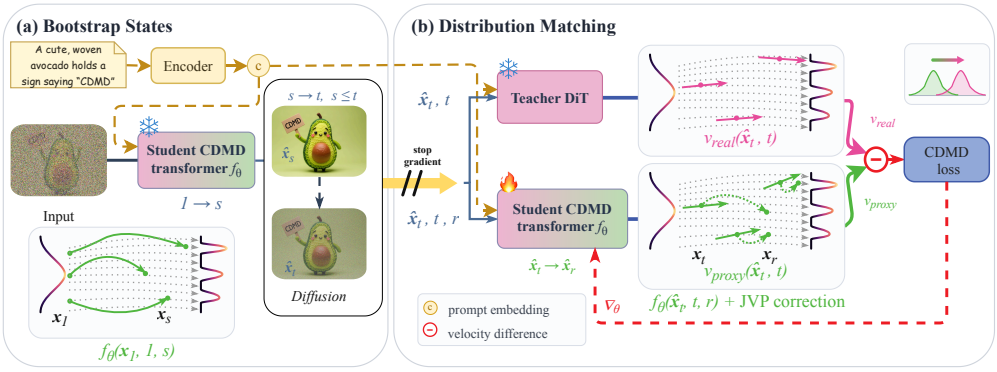


Figure 3: Overview of our Consistent Distribution Matching method. Panel (a) illustrates an effective data-free construction of $\hat{\boldsymbol{x}}_t$. Note that s is the balancing time between student transition and forward path. In (b), we demonstrate how the student proxy velocity matches the target teacher velocity. t, r are the start and end time, respectively. Note that the teacher model is the only supervision source. We include the architectural detail of our CDMD transformer in the Appendix.

training distribution to evolve with the student. At these states, we regress on the student velocity surrogate toward the teacher instantaneous velocity target through a single objective.

We further establish a theoretical connection between local transport consistency and distribution matching. Under appropriate assumptions, minimizing population loss implies that the distributions induced by the student flow map converge to the teacher marginals in squared Wasserstein-2 distance averaged over time. The minimizer of our objective becomes the teacher average velocity. Together, these results explain how local consistency can identify the teacher's transport behavior and achieve distributional alignment without auxiliary score networks.

Our contributions are threefold:

1. We introduce Consistent Distribution Matching Distillation, a data-free and simulation-free distillation framework for diffusion and flow-based generative models. The method uses only a frozen teacher and a trainable student, avoiding external distillation data, teacher rollouts and auxiliary proxy networks.
2. We provide a distribution-matching analysis showing that minimizing the proposed CDMD objective leads to narrowing the gap between the teacher distribution and the student distribution under regularity assumptions. The analysis also identifies the teacher's average velocity field as the unique global minimizer for our objective.
3. We analyze the variance of the JVP correction in the data-free setting and validate the analysis on synthetic datasets. These results clarify when the correction improves stability and when student-teacher path mismatch can degrade training. On ImageNet256×256, Consistent Distribution Matching Distillation achieves a 1-NFE FID of **2.05** and a 4-NFE FID of **1.37**, achieving state-of-the-art performance among data-free distillation methods. In the text-to-image comparison when training data is inaccessible, our method leads most preference and alignment metrics and FID.

2 Background

Diffusion and Flow Matching. Let p_0 denote the data distribution and p_1 a prior distribution. For each $\boldsymbol{x}_0 \sim p_0$ and $\boldsymbol{x}_1 \sim p_1$, Flow Matching (FM) (Albergo et al., 2023; Lipman et al., 2023; Liu et al., 2023) defines a continuous flow path between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ via an affine transformation $\boldsymbol{x}_t := \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{x}_1,\ t \in [0,1]$, where $\alpha_t, \sigma_t : [0,1] \to [0,1]$ are prescribed schedules. These schedules are smooth, monotone functions with the boundary conditions $\alpha_0 = \sigma_1 = 1, \alpha_1 = \sigma_0 = 0$. One standard approach constructs a path $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$ between a noise sample $\boldsymbol{x}_1$ and a data sample $\boldsymbol{x}_0$, which induces a constant conditional velocity (Lipman et al., 2023; Tong et al., 2023): $\boldsymbol{v}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{x}_1) = \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{x}_1 - \boldsymbol{x}_0$. By conditioning over all possible data and prior pairs

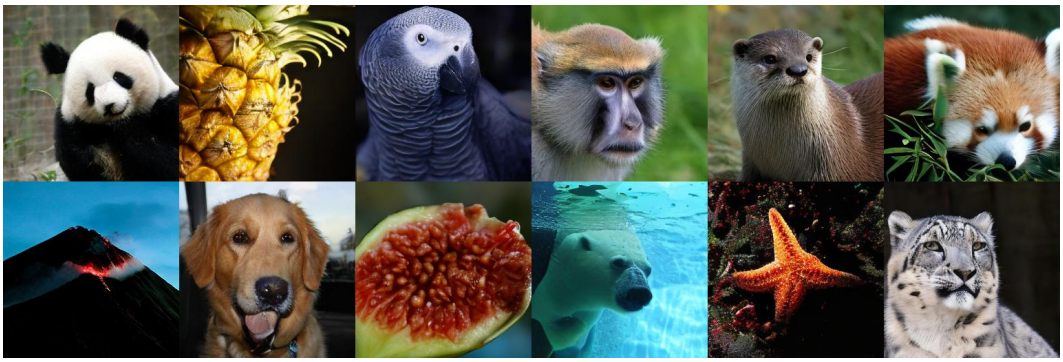


Figure 4: Uncurated qualitative Results on ImageNet 256×256 dataset. Samples are generated by our CDMD method with **one** NFE conditioned on the class labels.

$(\boldsymbol{x}_0, \boldsymbol{x}_1)$, we obtain a marginal probability path $\boldsymbol{x}_t \sim p_t$ and the corresponding target velocity field $\boldsymbol{v} : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$,

$$\boldsymbol{v}(\boldsymbol{x}_t, t) := \mathbb{E}_{p(\boldsymbol{x}_0, \boldsymbol{x}_1 | \boldsymbol{x}_t)}[\boldsymbol{v}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{x}_1)]. \tag{1}$$

Here $\boldsymbol{v}(\cdot, t)$ denotes the marginal velocity field at time t and Eq. (1) relates the marginal velocity field to the conditional velocity $\boldsymbol{v}_t$. Due to the intractability of the marginal velocity field $\boldsymbol{v}(\boldsymbol{x}_t, t)$, FM therefore learns a neural velocity field $\boldsymbol{v}_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$ by minimizing the conditional FM (CFM) objective

$$\mathcal{L}_{\mathrm{CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{v}_t, t}\left[\|\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{v}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{x}_1)\|_2^2\right], \tag{2}$$

for all $t \in [0,1]$. The minimizer $\boldsymbol{\theta}^\star$ of Eq. (2) provably recovers the true marginal velocity field at optimum, *i.e.*, $\boldsymbol{v}_{\boldsymbol{\theta}^\star}(\boldsymbol{x}_t, t) = \boldsymbol{v}(\boldsymbol{x}_t, t)$ (Lipman et al., 2023; Tong et al., 2023). During inference, we solve an ordinary differential equation (ODE) $\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{v}_{\boldsymbol{\theta}^\star}(\boldsymbol{x}_t, t)$ via integration.

Score Distillation. Score distillation aligns the student distribution q_0 with the teacher distribution p_0 by minimizing their reverse KL divergence, *i.e.* $D_{\mathrm{KL}}(q_0 \| p_0) = \mathbb{E}_{\hat{\boldsymbol{x}} \sim q_0}\left[\log q_0(\hat{\boldsymbol{x}}) - \log p_0(\hat{\boldsymbol{x}})\right]$. In practice, Distribution Matching Distillation (Yin et al., 2024a;b) (DMD) applies this objective to diffused distributions q_t and p_t. Given a generated sample $\hat{\boldsymbol{x}}_0 = G_{\boldsymbol{\theta}}(\boldsymbol{z})$ from the student, DMD constructs $\hat{\boldsymbol{x}}_t = \alpha_t \hat{\boldsymbol{x}}_0 + \sigma_t \boldsymbol{\varepsilon}$ and updates the generator using

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{SD}} = \mathbb{E}_t\left[\nabla_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q_t \| p_t)\right] \approx \mathbb{E}_{\boldsymbol{z}, t, \boldsymbol{\varepsilon}}\left[w_t \big(s_{\mathrm{fake}}(\hat{\boldsymbol{x}}_t, t) - s_{\mathrm{real}}(\hat{\boldsymbol{x}}_t, t)\big)^\top \left(\frac{\partial G_{\boldsymbol{\theta}}(\boldsymbol{z})}{\partial \boldsymbol{\theta}}\right)\right], \tag{3}$$

where $\boldsymbol{z}, \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$, and w_t is an adaptive reweighting term. The frozen teacher provides the real score $s_{\mathrm{real}} \approx \nabla_{\boldsymbol{x}} \log p_t$. In addition, an auxiliary network estimates the fake score $s_{\mathrm{fake}} \approx \nabla_{\boldsymbol{x}} \log q_t$ through standard diffusion training, using student generated clean samples as targets. Following variational score distillation (Wang et al., 2023), this procedure alternates between updating the generator and training the fake score model with separate objectives. Similar to GANs (Huang et al., 2024; Brock et al., 2018; Kang et al., 2023; Sauer et al., 2025; 2022), this coupled two-player optimization makes DMD intrinsically difficult to train. Under the two-timescale update rule, the fake score model is updated substantially more often than the generator, consuming a large fraction of the training compute to track the generator's evolving distribution, despite being discarded entirely at inference time. Although subsequent methods, including Score identity Distillation (SiD), FreeFlow, and SenseFlow (Zhou et al., 2024b;a; Tong et al., 2025; Ge et al., 2026), improve the optimization and training objective, they still follow the same paradigm by devoting substantial computation to training an auxiliary fake score model. In particular, SiD extends reverse KL to Fisher divergence, *i.e.* $D_{\mathrm{F}}(q_t \| p_t) = \mathbb{E}_{\hat{\boldsymbol{x}} \sim q_t} \|\nabla_{\hat{\boldsymbol{x}}} \log q_t(\hat{\boldsymbol{x}}) - \nabla_{\hat{\boldsymbol{x}}} \log p_t(\hat{\boldsymbol{x}})\|^2$.

3 Consistent Distribution Matching Distillation

This section develops Consistent Distribution Matching Distillation as a distillation method without external data or teacher simulation, as illustrated in Fig. 3. We first motivate supervision on states bootstrapped from the student model. Beyond the tradition KL divergence, we develop Consistent Distribution Matching from a Fisher-divergence criterion using a velocity surrogate derived from the student network. Finally, we relate the resulting consistency residual to distribution matching under regularity assumptions.

3.1 MINIMAL DESIGN ON DATA-FREE DISTILLATION

Current distillation methods require either online training with external data or expensive offline generation of teacher samples. In the former setting, training states are drawn from a forward FM path $\tilde{p}_t$, while the teacher follows its own probability path p_t. Since $\tilde{p}_t \neq p_t$ in general, the student is supervised on a mismatched state distribution. Training directly on p_t avoids this mismatch, but sampling intermediate teacher states requires simulating the teacher dynamics, introducing substantial computational overhead.

Unifying proxy model and student generator. Motivated by SiD (Zhou et al., 2024b), we therefore adopt a data-free construction based on the student rollout. However, evaluating this objective typically introduces an auxiliary model to estimate the student's marginal velocity, which violates our design principle of involving only the teacher and student. To eliminate this extra model, the student must be capable of both generating clean samples and providing a surrogate for its own marginal velocity within a unified framework.

We write the Fisher-divergence objective in velocity form under the FM framework[1]:

$$\mathcal{L}_{\mathrm{VD}}(\boldsymbol{\theta}) = \mathbb{E}_t \left[w_t D_{\mathrm{F}}(q_t \| p_t) \right] = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}} \left[\bar{w}_t \left\| \boldsymbol{v}_{\mathrm{fake}}(\hat{\boldsymbol{x}}_t, t) - \boldsymbol{v}_{\mathrm{real}}(\hat{\boldsymbol{x}}_t, t) \right\|_2^2 \right], \tag{4}$$

where $\bar{w}_t = (1-t)^2 t^{-2} w_t$ is a weighting function for $0 < t < 1$. Consequently, the problem reduces to jointly modeling global one-step generation and local instantaneous velocity. This calls for a model that maps between arbitrary times along an ODE trajectory by representing the integrated velocity, *i.e.* a flow map model (Albergo & Vanden-Eijnden, 2023; Boffi et al., 2024; Geng et al., 2025a;b; Zhang et al., 2026). We therefore opt to model the average velocity field. Define the average velocity $\boldsymbol{u} : \mathbb{R}^D \times [0,1] \times [0,1] \to \mathbb{R}^D$, for $0 \leq r < t \leq 1$, as $\boldsymbol{u}(\boldsymbol{x}_t, t, r) := \frac{\boldsymbol{x}_t - \boldsymbol{x}_r}{t-r} = \frac{1}{t-r} \int_r^t \boldsymbol{v}(\boldsymbol{x}_\tau, \tau)\, \mathrm{d}\tau$.

If r is independent of t, we recover the instantaneous velocity along the underlying ODE trajectory by differentiating this definition with respect to t. This identity motivates replacing the auxiliary proxy model with a velocity surrogate derived from the student itself. In particular, one-step generation and the velocity surrogate take the forms $G_{\boldsymbol{\theta}}(\boldsymbol{z}) = \boldsymbol{z} - \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)$, $\boldsymbol{v}_{\mathrm{fake}}(\boldsymbol{x}_t, t) \approx \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) + (t-r) \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$, respectively, where we unify the proxy model and the generator with $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$. For the surrogate, the total derivative is evaluated in the teacher's velocity direction, holding r fixed. Note that the average-velocity quotient is undefined at $r = t$. Hence, we define its continuous extension as $\boldsymbol{u}(\boldsymbol{x}_t, t, t) := \lim_{r \to t} \boldsymbol{u}(\boldsymbol{x}_t, t, r) = \boldsymbol{v}(\boldsymbol{x}_t, t)$, assuming the underlying velocity is continuous.

One advantage of modeling average velocity is that we can recover instantaneous velocity from its definition. This parameterization also allows us to generate a clean sample in one model pass. Thus, a single student network provides both one-step generation and instantaneous velocity modeling, eliminating the auxiliary network. Substituting the one-step generation and the velocity surrogate into Eq. (4) yields the following surrogate for the Fisher-divergence objective:

$$\mathcal{L}_{\mathrm{VD}}(\boldsymbol{\theta}) \approx \mathbb{E}_{t,r,\boldsymbol{z},\boldsymbol{\varepsilon}} \left[\bar{w}_t \left\| \left[\boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, r) + (t-r) \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, r) \right] - \boldsymbol{v}_{\mathrm{real}}(\hat{\boldsymbol{x}}_t, t) \right\|_2^2 \right]. \tag{5}$$

This substitution is exact when the velocity surrogate coincides with the marginal velocity of q_t.

Training objective. In the absence of external data, the central question is how to construct the intermediate noisy state $\boldsymbol{x}_t$. Our construction is motivated by two central desiderata. First, we aim to perform distillation without access to real training data, requiring the intermediate states to be generated from the student dynamic path. Second, to keep the procedure simulation-free, each intermediate state involves only a single evaluation of the student model. Let $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) := \boldsymbol{x}_t - (t-r)\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$ be the endpoint of the student transport. Leveraging the closed-form forward noising path, we therefore construct the target $\hat{\boldsymbol{x}}_t$ by first denoising the prior sample $\boldsymbol{z} \sim p_1$ to an earlier time s, followed by analytically renoising the resulting state from s to t:

$$\hat{\boldsymbol{x}}_t = \frac{1-t}{1-s} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, s) + \frac{t-s}{1-s} \boldsymbol{\varepsilon} \quad \boldsymbol{z}, \boldsymbol{\varepsilon} \sim p_1, \quad 0 \leq s \leq t < 1. \tag{6}$$

[1]We show the equivalence between the score and velocity formulations of reverse KL in the Appendix, allowing us to use the velocity formulation directly. The same applies for Fisher divergence. The precise form depends on the parameterization of the diffusion or flow-based model.

The two extremes of s recover familiar constructions. At $s = t$, the state lies on the student path, $\hat{\boldsymbol{x}}_t = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, t)$. At $s = 0$, it is a noised one-step student sample, $\hat{\boldsymbol{x}}_t = (1-t)\hat{\boldsymbol{x}}_0 + t\boldsymbol{\varepsilon}$, as in score distillation. We denote the resulting law of $\hat{\boldsymbol{x}}_t$ by $q_t^{\boldsymbol{\theta}}$ and refer to it as the student probability path. As the student approaches the teacher, $q_t^{\boldsymbol{\theta}}$ moves toward the teacher path and supplies more informative supervision states. We train the student with

$$\mathcal{L}_{\text{CDMD}}(\boldsymbol{\theta}) := \mathbb{E}_{t,r,\,\hat{\boldsymbol{x}}_t \overset{\text{sg}}{\sim} q_t^{\boldsymbol{\theta}}} \left\| \boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, r) + (t-r) \cdot \text{sg}\left[\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, r) \right] - \boldsymbol{v}(\hat{\boldsymbol{x}}_t, t) \right\|^2, \tag{7}$$

where sg is the stop-gradient operator, and we overload this notation $\overset{\text{sg}}{\sim}$ to indicate the sampled state is detached from the computation graph. The total derivative $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}_{\boldsymbol{\theta}} = D_{\hat{\boldsymbol{x}}_t}\boldsymbol{u}_{\boldsymbol{\theta}} \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t} + \partial_t \boldsymbol{u}_{\boldsymbol{\theta}} = \text{JVP}\left(\boldsymbol{u}_{\boldsymbol{\theta}}; \left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}, 1, 0\right)\right)$ is computed by one Jacobian-vector product (JVP) in the forward mode. We study the choice of the tangent $\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}$ in Sec. 3.2. Detaching the JVP prevents higher-order gradients through the total derivative, which stabilizes training. Our objective uses the consistency residual motivated by Eq. (5), aiming to reduce the Fisher divergence between student and teacher distributions.

Let $p_t := \left(\boldsymbol{f}(\cdot, 1, t)\right)_{\#} p_1$ be the teacher marginal, *i.e.*, the law of $\boldsymbol{f}(\boldsymbol{z}, 1, t)$ with $\boldsymbol{z} \sim p_1$. Let the residual of our CDMD objective be $\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) := \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) + (t-r) \cdot \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) - \boldsymbol{v}(\hat{\boldsymbol{x}}_t, t)$. Theorem 3.1 compares the teacher joint law P with the training law $\mathsf{Q}_{\boldsymbol{\theta}}$ of $(r, t, \hat{\boldsymbol{x}}_t)$. Part (a) shows that if P is contiguous with respect to $(\mathsf{Q}_{\boldsymbol{\theta}_n})$, then $\mathcal{L}_{\text{CDMD}}(\boldsymbol{\theta}_n) \to 0$ implies that the W_2 distance between the student pushforwards and the teacher marginals, averaged over r, also goes to zero. Part (b) shows that if P and $\mathsf{Q}_{\bar{\boldsymbol{\theta}}}$ are mutually absolutely continuous on denoising intervals, then every global minimizer recovers the teacher average velocity. Student-bootstrapped states are therefore a principled substitute for teacher rollouts, provided the student path stays close to the teacher path. The detailed theorem and proof are in Appendix A.

Theorem 3.1 (Distributional consistency and identification (simplified)). *Let p_t denote the teacher marginals and ν the reference-time law on $[0, 1)$. Under the respective regularity, coverage, and integrability assumptions detailed in Appendix A:*

- *(a) For any parameter sequence $(\boldsymbol{\theta}_n)$, vanishing population loss $\mathcal{L}_{CDMD}(\boldsymbol{\theta}_n) \to 0$ implies $\int_{[0,1)} W_2^2((\boldsymbol{f}_{\boldsymbol{\theta}_n}(\cdot, 1, r))_{\#} p_1,\, p_r)\, \nu(\mathrm{d}r) \to 0$.*
- *(b) If an admissible student can represent the teacher average velocity $\boldsymbol{u}$, the minimum population loss is zero, and every global minimizer recovers $\boldsymbol{u}$ almost surely under its training distribution.*

Sampling. Once $\boldsymbol{u}_{\boldsymbol{\theta}}$ is learned, the corresponding flow map is recovered by $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) = \boldsymbol{x}_t - (t-r)\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$. In particular, one-step generation over the full interval $(r, t) = (0, 1)$ is given by $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_1, 1, 0) = \boldsymbol{x}_1 - \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_1, 1, 0)$, where $\boldsymbol{x}_1 \sim p_1$. Few-step sampling can be obtained by applying the same map over a sequence of consecutive intervals. Therefore, this parameterization $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$ naturally enables inference-time scaling, as shown by the ablation in Fig. 6.

Tab. 4 shows that CDMD uniquely combines data-free and simulation-free distillation with only two models and a single optimizer, avoiding both adversarial training and the external real data required by GAN-based objectives.

3.2 Variance of the JVP correction

The detached JVP correction in $\mathcal{L}_{\text{CDMD}}$ is sensitive to the tangent direction used to evaluate it. In our notation, this direction is the input direction $(\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t, 1, 0)$ of the JVP, rather than the derivative vector $(\partial_{\boldsymbol{x}}\boldsymbol{u}_{\boldsymbol{\theta}}, \partial_t \boldsymbol{u}_{\boldsymbol{\theta}}, \partial_r \boldsymbol{u}_{\boldsymbol{\theta}})$ itself. Hence, in the data-free setting, where supervision is obtained only from the teacher on student-generated states, the choice of $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t$ directly affects the stochasticity and bias of the detached target. For any realization of $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t$, we define the corresponding JVP correction $\mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right) := \text{JVP}\left(\boldsymbol{u}_{\boldsymbol{\theta}}; \left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}, 1, 0\right)\right)$.

To diagnose the stochasticity of the correction under data-free supervision, we first define the cross-centered projected JVP second moment

$$\mathcal{M}_{\text{JVP}}^{q,p}\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) := \mathbb{E}_{\hat{\boldsymbol{x}}_t \sim q_t^{\boldsymbol{\theta}}} \left[\left\| \mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right) - \mathbb{E}_{\boldsymbol{x}_t \sim p_t}\left[\mathcal{G}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t, r; \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}\right)\right] \right\|^2 \right].$$

Table 1: **Data-free distillation on class-conditional ImageNet256×256 dataset. Left**: Comparison using LightningDiT-XL/2 as the teacher. **Right**: Comparison using SiT-XL/2 and SiT-XL/2+VA-REPA as teachers. All numbers are reported with CFG when applicable, and ×2 in NFE indicates that CFG doubles the number of function evaluations at inference time. † denotes results reproduced by us on ImageNet256×256, while ⋆ denotes methods adapted by us to the data-free setting. Bold and underlined values indicate the best and second-best results, respectively.

Method	NFE	FID↓	IS↑
Teacher Diffusion / Flow Models			
SiT-XL/2 (Ma et al., 2024)	250×2	2.06	277.5
SiT-XL/2+VA-REPA (Yang et al., 2026)	250×2	1.31	309.7
LightningDiT (Yao et al., 2025)	250×2	1.35	295.3
Data-free Distillation			
Teacher: LightningDiT-XL/2 (FID = 1.35)			
FreeFlow (Tong et al., 2025)	1	4.28	264.9
DMD2† (Yin et al., 2024a)	1	3.43	317.2
MFD† (Zhao et al., 2026)	1	6.67	281.6
MeanFlow⋆ (Geng et al., 2025a)	1	4.19	213.8
sCM⋆ (Lu & Song, 2024)	1	4.21	266.9
FACM⋆ (Peng et al., 2025)	1	2.15	303.8
	2	1.79	**319.6**
Ours	1	2.05	292.0
	2	<u>1.62</u>	<u>317.3</u>
	4	**1.37**	306.9

Method	NFE	FID↓	IS↑
Data-free Distillation			
Teacher: SiT-XL/2 (FID = 2.06)			
DMD2† (Yin et al., 2024a)	1	3.93	**273.2**
FreeFlow w.o. DMD	1	14.5	182.4
FreeFlow (Tong et al., 2025)	1	4.13	268.7
MeanFlow⋆ (Geng et al., 2025a)	1	4.07	251.5
Ours	1	<u>3.47</u>	251.4
	2	**2.87**	<u>271.0</u>
Teacher: SiT-XL/2+VA-REPA (FID = 1.31)			
DMD2† (Yin et al., 2024a)	1	3.86	239.1
FreeFlow w.o. DMD	1	15.4	171.3
FreeFlow (Tong et al., 2025)	1	3.72	266.4
MeanFlow⋆ (Geng et al., 2025a)	1	4.47	214.5
Ours	1	<u>2.93</u>	**286.4**
	2	**2.60**	<u>272.8</u>

This term measures the deviation of the JVP correction computed on student-generated states from the average correction on the target path. It captures both the intrinsic fluctuation of the correction under $q_t^{\boldsymbol{\theta}}$ and the distributional mean shift between $q_t^{\boldsymbol{\theta}}$ and p_t. Note the intrinsic JVP variance is recovered as the special case $\mathcal{V}_{\text{JVP}} = \mathcal{M}_{\text{JVP}}^{q,q}$, and the decomposition can be found in Appendix C. We evaluate these statistics for the term $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t$ with three viable options, *i.e.* $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t \in \{\boldsymbol{v}(\hat{\boldsymbol{x}}_t, t),\ \boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, t),\ \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)\}$. The first two terms are the teacher velocity and the student limiting velocity, while the last term is commonly known as the conditional velocity field $\boldsymbol{x}_1 - \hat{\boldsymbol{x}}_0$. We find that using $\boldsymbol{v}(\hat{\boldsymbol{x}}_t, t)$ works best, shown in Fig. 5 (c) the on synthetic dataset.

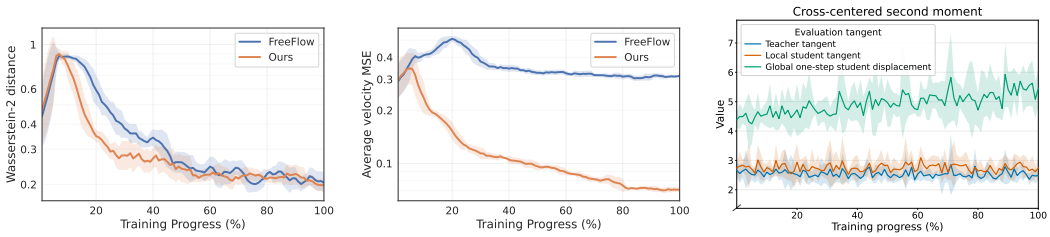
4 Experiments

Below we test CDMD on a synthetic 2D benchmark, class-conditional ImageNet 256×256, and text-to-image generation with SANA-1.6B. We then ablate guidance, the training-state distribution, and the number of sampling steps.

Setup. On ImageNet 256×256, we distill three teachers: SiT-XL/2 (Ma et al., 2024), SiT-XL/2 trained with VA-REPA (Yang et al., 2026), and LightningDiT-XL/2 with a VA-VAE autoencoder (Yao et al., 2025). The student uses the teacher architecture, with extra embeddings for r and w (Appendix E). We report FID-50K and IS. All baselines use the same training budget and the same evaluation pipeline. For text-to-image, we follow the protocol of MFD (Zhao et al., 2026) with a SANA-1.6B teacher (Xie et al., 2024a). Full hyperparameters are listed in Tab. 5.

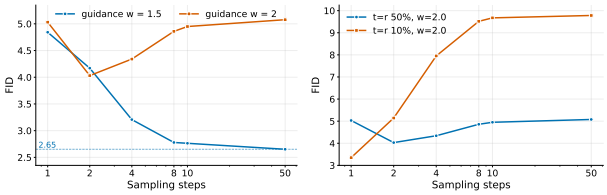
Synthetic benchmark. We first use a 2D spiral, where both the target distribution and the ground-truth average velocity are known. The exact velocity field serves as the teacher. This removes the effect of a learned teacher checkpoint and of image metrics. We track two quantities during training (Fig. 5). First, CDMD and FreeFlow reduce the endpoint W_2 distance at the same rate. Endpoint matching therefore cannot distinguish the two methods. Second, only CDMD keeps reducing the teacher average-velocity MSE, while FreeFlow plateaus at a much higher value. This suggests that *matching the endpoint marginal does not imply learning the teacher transport map*. The full metrics in Tab. 6 show the same pattern. Marginal Wasserstein scores are tied, but CDMD has lower coupled endpoint error and lower teacher average-velocity error. Appendix E provides the full setup. In Fig. 5 (c), we observe that JVP yields the most stable behavior when computed in the teacher direction, even on student-bootstrapped states.

ImageNet 256×256. Tab. 1 groups data-free methods by teacher under matched training budgets. CDMD achieves the best 1-NFE FID across all three teachers: 3.47 with SiT-XL/2, 2.93 with SiT-



(a) Endpoint W_2 distance. (b) Teacher average-velocity MSE. (c) Second moment of the JVP term.

Figure 5: **Endpoint matching does not imply transport matching.** Results on the 2D spiral dataset, where the teacher is the exact velocity field. (a) CDMD and FreeFlow reach similar endpoint distances to the target, showing comparable marginal matching ability. (b) CDMD learns the true teacher average velocity accurately, while FreeFlow has a higher error. (c) Among the three tangent choices in Sec. 3.2, the teacher tangent provides the lowest and most stable JVP second moment.



(a) (b)

Figure 6: **Multi-step gains depend on guidance and on the fraction of $t{=}r$ pairs.** (a) FID (↓) against sampling steps for two guidance scales. (b) FID (↓) against sampling steps for two $t{=}r$ ratios at $w = 2.0$.

m	Peak step	FID ↓	IS ↑
0	41K	3.62	251.5
0.25	43K	3.67	239.9
0.50	41K	**3.41**	254.8
0.75	44K	3.51	261.4
0.875	42K	3.57	**262.7**
1.00	44K	3.84	259.6

Table 2: **Mixing both noisy state types beats either alone.** 1-NFE results at the best checkpoint for each mixing ratio m, trained for 50K iterations.

XL/2+VA-REPA, and 2.05 with LightningDiT-XL/2. It also benefits more from stronger teachers: replacing SiT-XL/2 with VA-REPA improves CDMD by 0.54 FID, compared with 0.41 for FreeFlow and 0.07 for DMD2, while MeanFlow degrades. Without retraining, the same student further improves with additional NFEs, reaching FIDs of 2.60 with VA-REPA and 1.62 with LightningDiT-XL/2 at 2 NFEs, and 1.37 at 4 NFEs, close to the teacher's 1.35. We omit FreeFlow without DMD in the LightningDiT block because it fails to converge within the training budget. For data-free baselines marked with $\star$, including MeanFlow, we construct training states by noising clean samples from the current one-step student, corresponding to the $s = 0$ case of CDMD; these adaptations remain substantially worse.

Inference-time scaling. FreeFlow (Tong et al., 2025) is anchored to the prior. Its student only takes noise at $t = 1$, so it cannot be applied over sub-intervals. In contrast, CDMD learns $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$ for any interval, so extra steps are available at inference time. Fig. 6 (a) shows that the benefit depends on guidance. With $w = 1.5$, FID keeps improving with more steps and approaches the dashed reference. With $w = 2.0$, FID is best at very few steps and worsens as steps increase. This suggests that strong guidance helps one-step generation but amplifies the student–teacher mismatch over multi-step sampling. Fig. 12 in the appendix shows that the gain from extra steps also depends on the sampler.

Training-state construction. Alg. 1 samples $m \in [0, 1]$ to interpolate the two states in Eq. (6): larger m emphasizes one-step student samples, while smaller m emphasizes shared-noise interval states. Tab. 2 shows that intermediate values work best, with the lowest FID at $m = 0.5$ and highest IS at $m = 0.875$, while convergence speed remains similar. Fig. 6b shows a related trade-off for the fraction of $t = r$ pairs: 10% improves 1-NFE FID but degrades multi-step sampling, whereas 50% yields more stable few-step performance. This suggests that sufficient exposure to local intervals is important for multi-step sampling.

Text-to-image generation. We distill SANA-1.6B (Xie et al., 2024a), a flow-matching text-to-image model. The teacher and student are initialized from the same checkpoint. We follow the protocol of MFD (Zhao et al., 2026), training and evaluating on LAION-Aesthetics 6.5+ (Schuhmann et al., 2022). We report Aesthetic Score, PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023a), ImageReward (Xu et al., 2023), CLIPScore (Hessel et al., 2021), and FID against COCO-5k 2017.

Table 3: **CDMD leads on aesthetics, preference, and FID at 4 NFEs.** Text-to-image generation with SANA-1.6B. Baseline numbers follow (Zhao et al., 2026). Bold marks the best result among learned 4-NFE methods. † denotes methods reproduced by us. FID is computed against the COCO-5K 2017 validation set.

Model	NFE ↓	Aesthetic ↑	PS ↑	HPSv2 ↑	IR ↑	CLIPScore ↑	FID ↓
SANA teacher (Xie et al., 2024a)	20	6.335	0.2021	0.2417	0.8363	30.717	30.39
SANA teacher	4	5.976	0.1872	0.2173	-0.1008	28.251	35.65
SenseFlow (Ge et al., 2026)	4	6.482	0.1998	**0.2445**	0.7925	29.321	–
AYF (Sabour et al., 2025)	4	6.231	0.1963	0.2298	0.7022	29.176	–
Diff-Instruct (Luo et al., 2023b)	4	6.453	0.1985	0.2434	0.7677	29.393	–
Consistency Distillation† (Song et al., 2023)	4	6.248	0.1906	0.2263	0.6743	28.890	54.38
DMD2† (Yin et al., 2024a)	4	6.489	0.2006	0.2439	0.8042	29.103	32.34
MFD† (Zhao et al., 2026)	4	6.541	0.2002	0.2442	0.8483	**29.430**	28.77
CDMD (ours)	4	**6.906**	**0.2011**	0.2387	**0.8864**	28.992	**27.83**

Tab. 3 gives three findings. First, among learned 4-NFE methods, CDMD has the best Aesthetic Score, PickScore, and ImageReward. Its FID of 27.83 is also below MFD's 28.77. Second, CDMD beats the teacher run at 4 NFEs on all six metrics, and it beats the 20-NFE teacher on Aesthetic Score, ImageReward, and FID. Third, CDMD trails on HPSv2 and CLIPScore, where four and five of the six baselines score higher, respectively. These results show that CDMD works in one large-scale text-to-image setting. They do not establish statistical superiority. CDMD is also cheap to train. It uses about 7 GPU hours, compared with 22 for Diff-Instruct, 25 for MFD, and 30 for DMD2 (Tab. 7). Only consistency distillation is cheaper ($\sim$4 GPU hours), yet it scores lower on all five point metrics. All costs are measured on one NVIDIA RTX PRO 6000 GPU, the same setup as the baseline measurements of (Zhao et al., 2026).

5 Related Work

Trajectory Distillation. Distillation accelerates diffusion sampling by training students to reproduce the teacher's dynamics in fewer steps. Knowledge distillation (Zheng et al., 2023) learns a direct mapping from noise to data, while progressive distillation (Salimans & Ho, 2022; Meng et al., 2023) repeatedly halves the number of sampling steps. Consistency models (Lu & Song, 2024; Song & Dhariwal, 2023; Song et al., 2023; Luo et al., 2023a) map states along the same ODE trajectory to a shared clean sample. Flow map distillation (Boffi et al., 2025; Geng et al., 2025a;b; Sabour et al., 2025; Lee et al., 2025) extends this idea to transitions between arbitrary times along a trajectory. BOOT (Gu et al., 2023), π-Flow (Chen et al., 2025), and rectified flow (Liu et al., 2023) explore ways to reduce reliance on real training data. For methods that require teacher reference trajectories, generating them during training or precomputing them can limit scalability as tasks become more complex. Our method avoids numerically integrating the teacher during training. Compared with FreeFlow (Tong et al., 2025), which anchors its flow map at noise, our model supports transitions from arbitrary input states and times.

6 Conclusion

We presented CDMD, a data-free and simulation-free framework for distilling pretrained diffusion and flow-based teachers into efficient one-step or few-step flow map students. Our method replaces external distillation data, teacher rollouts, backward simulation, and auxiliary proxy networks with a student-induced training distribution and a single local average-velocity consistency objective. Theoretically, we showed that, under regularity assumptions, minimizing our CDMD objective leads to convergence of student sampling distributions to teacher distributions in Wasserstein distance, averaged over reference times, and every global minimizer recovers the teacher average velocity on sampled states. Empirically, CDMD achieves state-of-the-art ImageNet256$\times$256 generation quality among data-free distillation methods and further benefits from inference-time scaling through its parameterization. These results suggest that strong data-free distillation can be achieved with a simpler two-model training pipeline, while preserving the flexibility of multi-step sampling. Future work includes improving the stability of the JVP term, developing initialization strategies when no model weights are available, and extending the framework to large-scale generative tasks where data scarcity problem becomes significant.

AI USE STATEMENT

In this work, we use generative AI tools for the following tasks that require disclosure under the ICLR 2027 AI Policy for Authors. First, we use generative AI tools for formulating mathematical claims and assisting in the writing of proofs. Next, the AI model provided feedback on our experimental methodology in the form of a simulated review. This feedback included the suggestion of a controlled ablation that separates the training-state construction from the regression target. All interpreted numbers come from our own experiments. We have not used generative AI tools to implement methods, or generate, clean, or analyze data. We did not use generative AI to create synthetic datasets, translate text, or analyze qualitative data. Additionally, we use the model to polish our writing and create teaser figures for our paper.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The training procedure is given in Alg. 1, and the objective is given in Eq. (7). Complete proofs of the theoretical results, together with their assumptions, are in Appendix A. Architectural changes to the student, all training hyperparameters in Tab. 5, the evaluation protocol, and the synthetic benchmark setup are described in Appendix E. The text-to-image prompts are listed in Appendix I. We will release the codebase and the checkpoints upon publication of the paper.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Nicholas Boffi, Michael Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *Advances in Neural Information Processing Systems*, 38: 33346–33382, 2025.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching. *arXiv preprint arXiv:2406.07507*, 2024.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. *arXiv preprint arXiv:2510.14974*, 2025.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *ArXiv*, abs/2410.12557, 2024.

Yuxiang Fu, Qi Yan, Lele Wang, Ke Li, and Renjie Liao. Moflow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17282–17293, June 2025.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Senseflow: Scaling distribution matching for flow-based text-to-image distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bhdvdYJ8WZ.

Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. *arXiv preprint arXiv:2406.14548*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

Jiatao Gu, Tianrong Chen, David Berthelot, Huangjie Zheng, Yuyang Wang, Ruixiang Zhang, Laurent Dinh, Miguel Angel Bautista, Josh Susskind, and Shuangfei Zhai. Starflow: Scaling latent normalizing flows for high-resolution image synthesis. *arXiv preprint arXiv:2506.06276*, 2025.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18062–18071, 2025.

Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37: 44177–44215, 2024.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10124–10134, 2023.

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.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *ArXiv*, abs/2310.02279, 2023.

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.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11523–11532, 2022.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled meanflow: Turning flow models into flow maps for accelerated sampling. *arXiv preprint arXiv:2510.24474*, 2025.

Sangyun Lee, Yilun Xu, Tomas Geffner, Giulia Fanti, Karsten Kreis, Arash Vahdat, and Weili Nie. Truncated consistency models. *arXiv preprint arXiv:2410.14895*, 2024.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17416–17425, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Wenze Liu and Xiangyu Yue. Learning to integrate diffusion odes by averaging the derivatives. *arXiv preprint arXiv:2505.14502*, 2025.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv*, abs/2310.04378, 2023a.

Tianze Luo, Haotian Yuan, and Zhuang Liu. Soflow: Solution flow models for one-step generative modeling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cjb03GNqYw.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

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.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. Flow-anchored consistency models, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pzHuesCvcO.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv.org*, abs/2201.00273, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

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 and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Shangyuan Tong, Nanye Ma, Saining Xie, and Tommi Jaakkola. Flow map distillation without data. *arXiv preprint arXiv:2511.19428*, 2025.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Zike Wu, Pan Zhou, Kenji Kawaguchi, and Hanwang Zhang. Fast diffusion model. *arXiv preprint arXiv:2306.06991*, 2023b.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024a.

Sirui Xie, Zhisheng Xiao, Diederik Kingma, Tingbo Hou, Ying Nian Wu, Kevin P Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *Advances in Neural Information Processing Systems*, 37:45073–45104, 2024b.

Bicheng Xu*, Qi Yan*, Renjie Liao, Lele Wang, and Leonid Sigal. Joint generative modeling of scene graphs and images via diffusion models. *Transactions on Machine Learning Research (TMLR)*, 2025.

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.

Robin Yadav, Qi Yan, Guy Wolf, Avishek Joey Bose, and Renjie Liao. Retro synflow: Discrete flow matching for accurate and diverse single-step retrosynthesis. *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Qi Yan, Zhengyang Liang, Yang Song, Renjie Liao, and Lele Wang. Swingnn: Rethinking permutation invariance in diffusion models for graph generation. *arXiv preprint arXiv:2307.01646*, 2023.

Qi Yan, Brian Zhang, Yutong Zhang, Daniel Yang, Joshua White, Di Chen, Jiachao Liu, Langechuan Liu, Binnan Zhuang, Shaoshuai Shi, et al. Trajflow: Multi-modal motion prediction via flow matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems 2025*, 2025.

Donglin Yang, Yongxing Zhang, Xin Yu, Liang Hou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Renjie Liao. Stable velocity: A variance perspective on flow matching, 2026. URL https://arxiv.org/abs/2602.05435.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 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. *arXiv preprint arXiv:2405.14867*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18431–18441, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. Alphaflow: Understanding and improving meanflow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=adacb4JTIv.

Yongxing Zhang, Donglin Yang, and Renjie Liao. Symmetricdiffusers: Learning discrete diffusion on finite symmetric groups. In *International Conference on Learning Representations*, volume 2025, pp. 5989–6014, 2025.

An Zhao, Shengyuan Zhang, Zhongjian Sun, Yixiang Zhou, Zejian Li, Ling Yang, Tianrun Chen, and Lingyun Sun. Mean flow distillation: Robust and stable distillation for flow matching models. *arXiv preprint arXiv:2606.11155*, 2026.

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International conference on machine learning*, pp. 42390–42402. PMLR, 2023.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. *arXiv preprint arXiv:2503.07565*, 2025a.

Linqi Zhou, Mathias Parger, Ayaan Haque, and Jiaming Song. Terminal velocity matching. *arXiv preprint arXiv:2511.19797*, 2025b.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Long and short guidance in score identity distillation for one-step text-to-image generation. *ArXiv 2406.01561*, 2024a.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, 2024b.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Guided score identity distillation for data-free one-step text-to-image generation. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=HMVDiaWMwM.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *The Thirteenth International Conference on Learning Representations*, 2025d. URL https://openreview.net/forum?id=lS2SGfWizd.

A PROOFS

Theorem A.1 (Distributional consistency and identification)**.** *Assume the characteristic regularity conditions in Appendix A. Let p_t be the teacher marginal probability path, ν a probability law on $[0,1)$. Define $\mathsf{P}(\mathrm{d}r,\mathrm{d}t,\mathrm{d}\boldsymbol{x}) := \nu(\mathrm{d}r)\,\frac{\mathbf{1}_{\{r<t<1\}}}{1-r}\,\mathrm{d}t\,p_t(\mathrm{d}\boldsymbol{x})$, to be the joint state measure and let $\mathsf{Q}_{\boldsymbol{\theta}}$ be the training state measure of the student model on $(r,t,\hat{\boldsymbol{x}}_t)$.*

(a) Consider a sequence $(\boldsymbol{\theta}_n)$. Suppose P is contiguous with respect to $(\mathsf{Q}_{\boldsymbol{\theta}_n})$ and $\left(\|\mathcal{R}_{\boldsymbol{\theta}_n}\|^2\right)$ is uniformly integrable under P. Then $\mathcal{L}_{CDMD}\left(\boldsymbol{\theta}_n\right) = \mathbb{E}_{\mathsf{Q}_{\boldsymbol{\theta}_n}}\|\mathcal{R}_{\boldsymbol{\theta}_n}\|^2 \to 0$ implies

$$\int_{[0,1)} W_2^2((\boldsymbol{f}_{\boldsymbol{\theta}_n}(\cdot,1,r))_{\#}p_1,\, p_r)\,\nu(\mathrm{d}r) \to 0.$$

where

$$\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,r) := \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,r) + (t-r)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,r) - \boldsymbol{v}(\hat{\boldsymbol{x}}_t,t).$$

(b) Suppose an admissible student represents the teacher average velocity $\boldsymbol{u}$ as a differentiable field on the teacher-flow domain used by the sampler. For every global minimizer $\bar{\boldsymbol{\theta}}$, assume $\mathsf{P} \ll \mathsf{Q}_{\bar{\boldsymbol{\theta}}}, \mathsf{Q}_{\bar{\boldsymbol{\theta}}}\big|_{\{r<t\}} \ll \mathsf{P}$. Then the minimum population loss is zero, and every global minimizer satisfies $\boldsymbol{u}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x},t,r) = \boldsymbol{u}(\boldsymbol{x},t,r)$ and is unique $Q_{\bar{\boldsymbol{\theta}}}$-almost surely.

Characteristic regularity. Throughout parts (a) and (b), assume that the teacher flow is measurable and that, for p_1-almost every $\boldsymbol{x}_1$, the trajectory $t \mapsto \boldsymbol{x}_t$ is absolutely continuous on $[0,1]$ and satisfies $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = \boldsymbol{v}(\boldsymbol{x}_t,t)$ almost everywhere. For every admissible parameter $\boldsymbol{\theta}$, assume that $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t,r)$ is jointly measurable and continuously differentiable in $(\boldsymbol{x},t)$ for each fixed r and $t > r$. In addition, for $\nu \otimes p_1$-almost every $(r,\boldsymbol{x}_1)$, assume that

$$k_{\boldsymbol{\theta},r}(t) := (t-r)\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,r), \qquad t \in (r,1], \tag{8}$$

extends to an absolutely continuous function on $[r,1]$ with $k_{\boldsymbol{\theta},r}(r) = 0$. These assumptions apply to every $\boldsymbol{\theta}_n$ in part (a) and every admissible field considered in part (b). We also assume $p_r \in \mathcal{P}_2(\mathbb{R}^D)$ for ν-almost every r.

Proof. (a) Let $\boldsymbol{x}_1 \sim p_1$ and let $\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_1,1,t)$ denote the corresponding teacher trajectory. Then $\boldsymbol{x}_t \sim p_t$ and $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = \boldsymbol{v}(\boldsymbol{x}_t,t)$ for almost every t. Throughout the proof, r is held fixed when differentiating in t. More precisely, the residual is the function

$$\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{x},t,r) := \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t,r) - \boldsymbol{v}(\boldsymbol{x},t) + (t-r)\left(\partial_t\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t,r) + D_{\boldsymbol{x}}\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t,r)\boldsymbol{v}(\boldsymbol{x},t)\right).$$

Along a teacher trajectory, the expression in parentheses equals $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,r)$. At a student-sampled state $\hat{\boldsymbol{x}}_t$, the same residual is evaluated at $\boldsymbol{x} = \hat{\boldsymbol{x}}_t$; the derivative still uses the teacher velocity $\boldsymbol{v}(\hat{\boldsymbol{x}}_t,t)$.

Let ν be the reference-time probability law on $[0,1)$, and define the teacher joint measure by

$$\mathsf{P}(\mathrm{d}r,\mathrm{d}t,\mathrm{d}\boldsymbol{x}) := \nu(\mathrm{d}r)\,\frac{\mathbf{1}_{\{r<t<1\}}}{1-r}\,\mathrm{d}t\,p_t(\mathrm{d}\boldsymbol{x}).$$

Let $\mathsf{Q}_{\boldsymbol{\theta}_n}$ denote the actual joint sampling law of $(r,t,\hat{\boldsymbol{x}}_t)$ used in the population loss, and write

$$G_n(r,t,\boldsymbol{x}) := \left\|\mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x},t,r)\right\|^2, \qquad \mathcal{L}_{\mathrm{CDMD}}\left(\boldsymbol{\theta}_n\right) = \mathbb{E}_{\mathsf{Q}_{\boldsymbol{\theta}_n}}G_n.$$

If the sampler includes $r = t$, the residual there is defined directly as $\mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{x},t,t) = \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t,t) - \boldsymbol{v}(\boldsymbol{x},t)$.

We first relate the residual along teacher trajectories to the transport error. For $\nu \otimes p_1$-almost every $(r,\boldsymbol{x}_1)$, define

$$\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t,t,r) := \boldsymbol{u}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t,t,r) - \boldsymbol{u}(\boldsymbol{x}_t,t,r).$$

The teacher average velocity satisfies

$$(t-r)\boldsymbol{u}(\boldsymbol{x}_t,t,r) = \boldsymbol{x}_t - \boldsymbol{x}_r.$$

Differentiating this identity and subtracting it from the student residual gives, for almost every $t \in (r, 1)$,

$$\begin{aligned}\mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r) &= \Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r) + (t-r)\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[(t-r)\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\right].\end{aligned}$$

Since

$$(t-r)\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r) = k_{\boldsymbol{\theta}_n, r}(t) - (\boldsymbol{x}_t - \boldsymbol{x}_r),$$

the characteristic assumptions imply that this function extends absolutely continuously to $[r, 1]$ with value zero at $t = r$. The fundamental theorem of calculus therefore gives

$$(1-r)\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) = \int_r^1 \mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\,\mathrm{d}t.$$

Using the definitions of the student and teacher maps, we obtain

$$\begin{aligned}\boldsymbol{f}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) - \boldsymbol{f}(\boldsymbol{x}_1, 1, r) &= -(1-r)\Delta_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) \\ &= -\int_r^1 \mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\,\mathrm{d}t.\end{aligned}$$

By Jensen's inequality for the uniform probability measure on $[r, 1]$,

$$\left\|\boldsymbol{f}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) - \boldsymbol{f}(\boldsymbol{x}_1, 1, r)\right\|^2 \le (1-r)\int_r^1 \left\|\mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\right\|^2 \mathrm{d}t.$$

Note that $\boldsymbol{f}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) \sim (\boldsymbol{f}_{\boldsymbol{\theta}_n}(\cdot, 1, r))_{\#}p_1$ and $\boldsymbol{f}(\boldsymbol{x}_1, 1, r) = \boldsymbol{x}_r \sim p_r$. Coupling these outputs through the same $\boldsymbol{x}_1$, and using Tonelli's theorem, yields

$$\begin{aligned}&W_2^2((\boldsymbol{f}_{\boldsymbol{\theta}_n}(\cdot, 1, r))_{\#}p_1, p_r) \\ &\quad\le \mathbb{E}_{\boldsymbol{x}_1 \sim p_1}\left\|\boldsymbol{f}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_1, 1, r) - \boldsymbol{f}(\boldsymbol{x}_1, 1, r)\right\|^2 \\ &\quad\le (1-r)\int_r^1 \mathbb{E}_{\boldsymbol{x}_t \sim p_t}\left\|\mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\right\|^2 \mathrm{d}t.\end{aligned}$$

Uniform integrability implies $\mathbb{E}_{\mathsf{P}}G_n < \infty$, so the last bound is finite for ν-almost every r. Together with $p_r \in \mathcal{P}_2(\mathbb{R}^D)$, this also ensures that the student pushforward has finite second moment for such r. Integrating over r gives

$$\begin{aligned}&\int_{[0,1)} W_2^2((\boldsymbol{f}_{\boldsymbol{\theta}_n}(\cdot, 1, r))_{\#}p_1, p_r)\,\nu(\mathrm{d}r) \\ &\quad\le \int_{[0,1)}(1-r)\int_r^1 \mathbb{E}_{\boldsymbol{x}_t \sim p_t}\left\|\mathcal{R}_{\boldsymbol{\theta}_n}(\boldsymbol{x}_t, t, r)\right\|^2 \mathrm{d}t\,\nu(\mathrm{d}r) \\ &\quad= \mathbb{E}_{\mathsf{P}}\left[(1-r)^2 G_n\right] \le \mathbb{E}_{\mathsf{P}}G_n. \qquad (9)\end{aligned}$$

It remains to show that $\mathbb{E}_{\mathsf{P}}G_n \to 0$. For every $\varepsilon > 0$, Markov's inequality gives

$$\mathsf{Q}_{\boldsymbol{\theta}_n}(G_n > \varepsilon) \le \frac{\mathcal{L}_{\text{CDMD}}(\boldsymbol{\theta}_n)}{\varepsilon} \longrightarrow 0.$$

By contiguity, applied to the measurable sets $\{G_n > \varepsilon\}$, we also have $\mathsf{P}(G_n > \varepsilon) \to 0$. For every $M > \varepsilon$, splitting the expectation gives

$$\mathbb{E}_{\mathsf{P}}G_n \le \varepsilon + M\,\mathsf{P}(G_n > \varepsilon) + \mathbb{E}_{\mathsf{P}}\left[G_n \mathbf{1}_{\{G_n > M\}}\right].$$

Taking $\limsup_{n\to\infty}$, then letting $M \to \infty$ using uniform integrability, and finally letting $\varepsilon \downarrow 0$, proves $\mathbb{E}_{\mathsf{P}}G_n \to 0$. Substituting this convergence into Eq. (9) proves the desired result.

(b) In this part, we use the same residual $\mathcal{R}_{\boldsymbol{\theta}}$ as in part (a). Since stop-gradient operators leave its numerical value unchanged, our population loss is

$$\mathcal{L}_{\text{CDMD}}(\boldsymbol{\theta}) = \mathbb{E}_{(r,t,\hat{\boldsymbol{x}}_t)\sim \mathsf{Q}_{\boldsymbol{\theta}}} \left\| \mathcal{R}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, r) \right\|^2. \tag{10}$$

We first show that zero residual and the characteristic boundary condition uniquely determine the field values along a teacher trajectory. Fix an admissible field $\boldsymbol{u}^*$, and choose $(r, \boldsymbol{x}_1)$ from the full-measure set on which its characteristic regularity conditions hold. Let $\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_1, 1, t)$ denote the corresponding teacher trajectory. Suppose that $\boldsymbol{u}^*$ has zero residual for almost every $t \in (r, 1)$. Then

$$\begin{aligned} 0 &= \boldsymbol{u}^*(\boldsymbol{x}_t, t, r) + (t-r)\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}^*(\boldsymbol{x}_t, t, r) - \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[(t-r)\boldsymbol{u}^*(\boldsymbol{x}_t, t, r) - (\boldsymbol{x}_t - \boldsymbol{x}_r)\right] \end{aligned}$$

for almost every $t \in (r, 1)$. By characteristic regularity, the expression in brackets is absolutely continuous on $[r, 1]$. Therefore,

$$(t-r)\boldsymbol{u}^*(\boldsymbol{x}_t, t, r) - (\boldsymbol{x}_t - \boldsymbol{x}_r) = C, \qquad r < t \le 1, \tag{11}$$

where $C \in \mathbb{R}^D$ is independent of t. We then show that $C = 0$. The characteristic boundary condition and continuity of the teacher trajectory give

$$\lim_{t\downarrow r}(t-r)\boldsymbol{u}^*(\boldsymbol{x}_t, t, r) = 0, \qquad \lim_{t\downarrow r}(\boldsymbol{x}_t - \boldsymbol{x}_r) = 0. \tag{12}$$

Taking $t \downarrow r$ in Eq. (11) therefore gives $C = 0$. It follows that

$$\boldsymbol{u}^*(\boldsymbol{x}_t, t, r) = \frac{\boldsymbol{x}_t - \boldsymbol{x}_r}{t-r} = \boldsymbol{u}(\boldsymbol{x}_t, t, r), \qquad r < t \le 1. \tag{13}$$

Here we used the teacher identity $(t-r)\boldsymbol{u}(\boldsymbol{x}_t, t, r) = \boldsymbol{x}_t - \boldsymbol{x}_r$.

We now consider the population loss. By the realizability assumption in part (b), there exists an admissible parameter $\boldsymbol{\theta}^{\mathrm{T}}$ representing the teacher average velocity and satisfying

$$\mathcal{R}_{\boldsymbol{\theta}^{\mathrm{T}}} = 0 \qquad \mathsf{Q}_{\boldsymbol{\theta}^{\mathrm{T}}}\text{-almost surely.} \tag{14}$$

Consequently, $\mathcal{L}_{\text{CDMD}}(\boldsymbol{\theta}^{\mathrm{T}}) = 0$. Since the loss is nonnegative, its global minimum is zero and $\boldsymbol{\theta}^{\mathrm{T}}$ is a global minimizer.

Now suppose $\bar{\boldsymbol{\theta}}$ is another global minimizer. We have

$$0 = \mathcal{L}_{\text{CDMD}}(\bar{\boldsymbol{\theta}}) = \mathbb{E}_{\mathsf{Q}_{\bar{\boldsymbol{\theta}}}} \left\| \mathcal{R}_{\bar{\boldsymbol{\theta}}} \right\|^2. \tag{15}$$

Nonnegativity implies that $\mathcal{R}_{\bar{\boldsymbol{\theta}}} = 0$ $\mathsf{Q}_{\bar{\boldsymbol{\theta}}}$-almost surely. By the coverage assumption $\mathsf{P} \ll \mathsf{Q}_{\bar{\boldsymbol{\theta}}}$, the residual also vanishes P-almost surely. The definition of P and Tonelli's theorem therefore give

$$\mathbb{E}_{\boldsymbol{x}_1\sim p_1} \int_r^1 \left\| \mathcal{R}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_t, t, r) \right\|^2 \mathrm{d}t = 0 \qquad \text{for } \nu\text{-almost every } r. \tag{16}$$

Hence, for $\nu \otimes p_1$-almost every $(r, \boldsymbol{x}_1)$, the residual vanishes for almost every $t \in (r, 1)$. Applying the uniqueness argument above with $\boldsymbol{u}^* = \boldsymbol{u}_{\bar{\boldsymbol{\theta}}}$ yields $\boldsymbol{u}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}_t, t, r) = \boldsymbol{u}(\boldsymbol{x}_t, t, r)$ for every $r < t \le 1$ along these trajectories, and consequently $\boldsymbol{u}_{\bar{\boldsymbol{\theta}}} = \boldsymbol{u}$ P-almost surely.

Finally, the reverse coverage assumption $\mathsf{Q}_{\bar{\boldsymbol{\theta}}}|_{\{r<t\}} \ll \mathsf{P}$ transfers this equality to the off-diagonal training law. On any sampled diagonal, zero residual gives $\boldsymbol{u}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, t, t) = \boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{u}(\boldsymbol{x}, t, t)$ directly. Therefore, the teacher average velocity is realized by a global minimizer, and every global minimizer satisfies

$$\boldsymbol{u}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}, t, r) = \boldsymbol{u}(\boldsymbol{x}, t, r) \qquad \mathsf{Q}_{\bar{\boldsymbol{\theta}}}\text{-almost surely,} \tag{17}$$

as desired.

□

B Additional Baseline

FreeFlow (Tong et al., 2025) observed that many existing distillation methods can suffer from *Teacher-Data Mismatch*: the student is trained on intermediate states sampled from data-noising distributions $\tilde{p}_t$, whereas the teacher follows its own generative distributions

$$\hat{p}_t := \big(\boldsymbol{f}(\cdot, 1, t)\big)_{\#} p_1, \tag{18}$$

where $\big(\boldsymbol{f}(\cdot, 1, t)\big)_{\#} p_1$ denotes the pushforward of p_1 under the teacher flow map $\boldsymbol{z} \mapsto \boldsymbol{f}(\boldsymbol{z}, 1, t)$; equivalently, if $\boldsymbol{z} \sim p_1$, then $\boldsymbol{f}(\boldsymbol{z}, 1, t) \sim \hat{p}_t$. This mismatch becomes problematic when the teacher distribution differs from the available data distribution, *e.g.*, due to generalization, guidance, post-training, or unavailable proprietary training data. FreeFlow (Tong et al., 2025) circumvents this issue by sampling only from the prior, since the noising process and the teacher generative process are guaranteed to coincide at $t = 1$, *i.e.*, $\tilde{p}_1 = \hat{p}_1 = p_1$.

In particular, FreeFlow (Tong et al., 2025) adopts a prior-anchored framework, which can be viewed as the converse of consistency models (Lu & Song, 2024; Song et al., 2023; Luo et al., 2023a). Let $\boldsymbol{z} \sim p_1$ and define the target time $r = 1 - \delta$, where $\delta \in [0, 1]$ is the duration of a single reverse-time transition from the prior time 1 to the target time r. Intuitively, FreeFlow learns to predict intermediate states directly from noise, rather than matching intermediate states obtained from the data-noising process. Using our notation, $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r)$ denotes the average velocity associated with this prior-anchored transition, with $\delta = 1 - r$. The student flow map from time 1 to time r is written as

$$\hat{\boldsymbol{x}}_r = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r) = \boldsymbol{z} - (1 - r)\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r), \qquad r \in [0, 1]. \tag{19}$$

Differentiating with respect to the jump duration δ gives

$$\partial_\delta \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 1 - \delta) = \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r) - (1 - r)\, \partial_r \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r), \qquad r = 1 - \delta. \tag{20}$$

Matching this generating velocity to the teacher velocity evaluated at the student-predicted state yields the FreeFlow prediction objective with only teacher and student involved

$$\mathcal{L}_{\text{FreeFlow w.o. DMD}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim p_1, r}\left[\left\|\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r) + \mathrm{sg}(\boldsymbol{v}(\hat{\boldsymbol{x}}_r, r) - (1 - r)\, \partial_r \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, r))\right\|_2^2\right], \tag{21}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. This stems from Eq. (26) by setting $t = 1$.

Since this prediction objective is self-referential, approximation errors in $\boldsymbol{f}_{\boldsymbol{\theta}}$ may accumulate along the generated trajectory. FreeFlow therefore introduces an additional correction term that aligns the noising velocity induced by the student-generated distribution with the teacher velocity. This correction is closely related to the variational score distillation objective (Wang et al., 2023). If $\boldsymbol{v}_{\mathrm{N}}(\hat{\boldsymbol{x}}_r, r)$ denotes the marginal noising velocity induced by the student-generated distribution $\hat{p}_0$, the correction signal aligns $\boldsymbol{v}_{\mathrm{N}}$ with the teacher velocity through the surrogate gradient

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{FreeFlow-VSD}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{z}, \boldsymbol{\varepsilon}, r}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)^\top \mathrm{sg}(\boldsymbol{v}_{\mathrm{N}}(\hat{\boldsymbol{x}}_r, r) - \boldsymbol{v}(\hat{\boldsymbol{x}}_r, r))\right], \tag{22}$$

where $\hat{\boldsymbol{x}}_r$ is obtained by first generating $\hat{\boldsymbol{x}}_0 = \boldsymbol{z} - \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)$ with $\boldsymbol{z} \sim p_1$ and then applying the noising process,

$$\hat{\boldsymbol{x}}_r = (1 - r) \cdot \hat{\boldsymbol{x}}_0 + r \cdot \boldsymbol{\varepsilon} = (1 - r)\big(\boldsymbol{z} - \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)\big) + r\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim p_1, \quad r \in [0, 1]. \tag{23}$$

In practice, $\boldsymbol{v}_{\mathrm{N}}$ is estimated by an auxiliary proxy FM model trained on student-generated samples, using either fine-tuning or LoRA adaptation. FreeFlow thus combines a trajectory-consistency prediction term with a distribution-level correction term, while requiring only prior samples from p_1. This prior-anchored formulation introduces a trade-off: it simplifies distillation by avoiding supervision over arbitrary timestep pairs, but confines training to trajectories initialized at $t = 1$. As a result, the student is optimized primarily as a one-step noise-to-data sampler, rather than as a general flow map capable of modeling denoising transitions between arbitrary timesteps (t, r) with $t > r$. Consequently, it reduces training complexity at the cost of limiting inference-time scaling and not naturally preserving the multi-step scaling behavior of the original flow model.

C MORE ON THE VARIANCE OF THE TARGET

The first uses the teacher tangent, the second uses the local student tangent, and the third uses the global one-step student displacement.

In practice, for a minibatch $\{\hat{\boldsymbol{x}}_t^{(i)}, t^{(i)}, r^{(i)}\}_{i=1}^B$, we estimate

$$\widehat{\mathcal{V}}_{\mathrm{JVP}} = \frac{1}{B}\sum_{i=1}^{B}\left\|\mathcal{G}_{\boldsymbol{\theta}}^{(i)} - \bar{\mathcal{G}}_{\boldsymbol{\theta}}\right\|^2, \qquad \bar{\mathcal{G}}_{\boldsymbol{\theta}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{G}_{\boldsymbol{\theta}}^{(i)},$$

where

$$\mathcal{G}_{\boldsymbol{\theta}}^{(i)} = \mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t^{(i)}, t^{(i)}, r^{(i)}; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t^{(i)}}{\mathrm{d}t}\right).$$

Similarly, the ground-truth projected error is

$$\widehat{\mathcal{D}}_{\mathrm{JVP}} = \frac{1}{B}\sum_{i=1}^{B}\left\|\mathcal{G}_{\boldsymbol{\theta}}^{(i)} - \mathcal{G}_{\boldsymbol{\theta},\mathrm{T}}^{(i)}\right\|^2,$$

where

$$\mathcal{G}_{\boldsymbol{\theta},\mathrm{T}}^{(i)} = \mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t^{(i)}, t^{(i)}, r^{(i)}; \boldsymbol{v}(\hat{\boldsymbol{x}}_t^{(i)}, t^{(i)})\right).$$

The projected variance measures the stability of the detached correction, while the ground-truth projected error measures its deviation from the teacher-consistent tangent.

To diagnose whether the JVP correction on student-generated states is aligned with the correction on the target path, we also consider a cross-centered projected JVP quantity. For a fixed tangent realization $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t$, define

$$\mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right) := \mathrm{JVP}\left(\boldsymbol{u}_{\boldsymbol{\theta}}; \left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}, 1, 0\right)\right).$$

Let

$$\mu_q\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) := \mathbb{E}_{\hat{\boldsymbol{x}}_t \overset{\mathrm{sg}}{\sim} q_t^{\boldsymbol{\theta}}}\left[\mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right)\right]$$

and

$$\mu_p\left(\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}; t, r\right) := \mathbb{E}_{\boldsymbol{x}_t \sim p_t}\left[\mathcal{G}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t, r; \frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}\right)\right].$$

Here, μ_q is the mean JVP correction on student-generated states, while μ_p is the mean JVP correction on the target path. We define the cross-centered projected JVP second moment as

$$\mathcal{M}_{\mathrm{JVP}}^{q,p}\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) := \mathbb{E}_{\hat{\boldsymbol{x}}_t \overset{\mathrm{sg}}{\sim} q_t^{\boldsymbol{\theta}}}\left[\left\|\mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right) - \mu_p\left(\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}; t, r\right)\right\|^2\right].$$

This quantity is not a variance under a single distribution. Instead, it decomposes as

$$\mathcal{M}_{\mathrm{JVP}}^{q,p}\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) = \mathcal{V}_{\mathrm{JVP}}^{q}\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) + \left\|\mu_q\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) - \mu_p\left(\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t}; t, r\right)\right\|^2,$$

where

$$\mathcal{V}_{\mathrm{JVP}}^{q}\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right) := \mathbb{E}_{\hat{\boldsymbol{x}}_t \overset{\mathrm{sg}}{\sim} q_t^{\boldsymbol{\theta}}}\left[\left\|\mathcal{G}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_t, t, r; \frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}\right) - \mu_q\left(\frac{\mathrm{d}\hat{\boldsymbol{x}}_t}{\mathrm{d}t}; t, r\right)\right\|^2\right].$$

Therefore, $\mathcal{M}_{\mathrm{JVP}}^{q,p}$ measures both the intrinsic stochasticity of the JVP correction on student-generated states and the mismatch between the average correction under $q_t^{\boldsymbol{\theta}}$ and the average correction under the target path p_t.

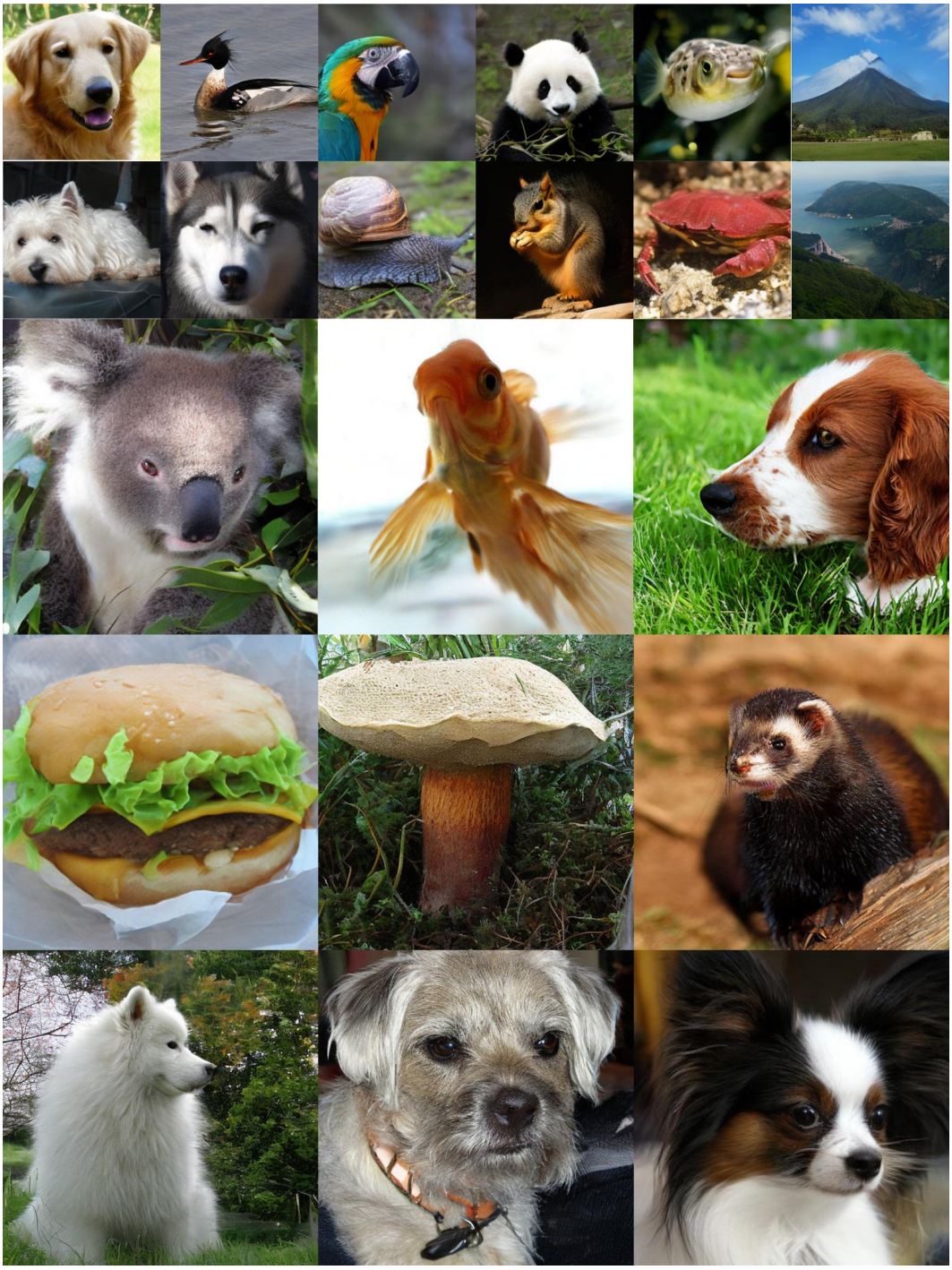


Figure 7: More qualitative results on ImageNet 256×256 and 512×512 datasets with 1-NFE sampling.

Table 4: Comparison of representative distillation methods. Some methods that can be trained from scratch, we consider their distillation variants, as our focus is exclusively on the distillation task. Among the compared methods, our method is the only approach that is simultaneously data-free and simulation-free while requiring only two models and a single optimizer. † indicates that the property may vary depending on the training setup. ⋆ indicates that these models can also be trained from scratch, whereas we only consider the distillation setting here.

Method	# Models	# Opts	Data-Free	Simulation-Free	GAN
BOOT (Gu et al., 2023)	2	1	✓	✗	✗
DMD2 (Yin et al., 2024a)	3	2	✓†	✓†	✓†
SiD-LSG (Zhou et al., 2025c)	3	2	✓	✓	✗
SiDA (Zhou et al., 2025d)	3	2	✗	✓	✓
FreeFlow (Tong et al., 2025)	3	2	✓	✓	✗
CTM (Kim et al., 2023)	3	2	✗	✓	✓
FACM (Peng et al., 2025)	2	1	✗	✓	✗
MF(Geng et al., 2025a)/CM(Lu & Song, 2024; Song et al., 2023)/IMM(Zhou et al., 2025a)	2⋆	1	✗	✓	✗†
CDMD (Ours)	**2**	**1**	✓	✓	✗

D Relationship with Prior Methods

In the notation of FreeFlow, $\boldsymbol{F_\theta}$ denotes the forward-time velocity, whereas our $\boldsymbol{u}(\cdot)$ denotes the corresponding backward-time velocity. Therefore, the two are related by $\boldsymbol{F_\theta} = -\boldsymbol{u}(\cdot)$. Beyond this notational difference, FreeFlow relies on pseudo-gradients propagated through student-generated trajectories, typically over intervals such as $[0, 1]$ or $[t, 1]$. Consequently, its updates depend on the quality of the intermediate student states $\hat{\boldsymbol{x}}_t$, which may introduce additional sensitivity during training. In contrast, our method avoids this rollout dependence by providing supervision directly on arbitrary local intervals $[r, t]$. This enables training across the denoising process without explicitly simulating intermediate student trajectories.

D.1 Flow maps

For a ground-truth velocity field $\boldsymbol{v}(\boldsymbol{x}_t, t)$, there exists a flow map (Albergo & Vanden-Eijnden, 2023; Boffi et al., 2024) $\boldsymbol{f} : \mathbb{R}^D \times [0, 1] \times [0, 1] \to \mathbb{R}^D$ that maps a state at time $t \in [0, 1]$ to its corresponding state at time $r \in [0, 1]$. Throughout this work, we primarily consider the backward map, *i.e.* a denoiser, with $0 \le r \le t \le 1$ for simplicity. More precisely, the flow map defines the ODE trajectory following $\boldsymbol{v}(\boldsymbol{x}_\tau, \tau)$ for all $\tau \in [r, t]$, *i.e.*,

$$\boldsymbol{f}(\boldsymbol{x}_t, t, r) := \boldsymbol{x}_t + \int_t^r \boldsymbol{v}(\boldsymbol{x}_\tau, \tau)\, \mathrm{d}\tau\,. \tag{24}$$

When $r < t$, the integral is taken in the backward direction and $\boldsymbol{f}(\boldsymbol{x}_t, t, r)$ returns the state at the earlier time r. In practice, sampling is carried out by numerically integrating the learned ODE using standard solvers such as Euler's method.

By viewing the marginal $\boldsymbol{v}(\boldsymbol{x}_t, t)$ as the instantaneous velocity field, MeanFlow and its variants (Geng et al., 2025a;b; Zhang et al., 2026) construct an average velocity field between times $0 \le r < t \le 1$ as

$$\boldsymbol{u}(\boldsymbol{x}_t, t, r) := \frac{\boldsymbol{x}_t - \boldsymbol{f}(\boldsymbol{x}_t, t, r)}{t - r} = \frac{1}{t - r} \int_r^t \boldsymbol{v}(\boldsymbol{x}_\tau, \tau)\, \mathrm{d}\tau\,,$$

where $\boldsymbol{f}(\boldsymbol{x}_t, t, r)$ denotes the flow map from time t to time r so that these one-step models can be trained from scratch. Equivalently, we have

$$\int_r^t \boldsymbol{v}(\boldsymbol{x}_\tau, \tau)\, \mathrm{d}\tau = (t - r) \cdot \boldsymbol{u}(\boldsymbol{x}_t, t, r). \tag{25}$$

Note that the flow map can also be recovered from the average velocity field via $\boldsymbol{f}(\boldsymbol{x}_t, t, r) = \boldsymbol{x}_t - (t - r)\boldsymbol{u}(\boldsymbol{x}_t, t, r)$. We define $\boldsymbol{u}(\boldsymbol{x}_t, t, t) := \lim_{r \to t} \boldsymbol{u}(\boldsymbol{x}_t, t, r) = \boldsymbol{v}(\boldsymbol{x}_t, t)$, where we assume

the limit exists and is equal to the velocity field at time t. Now, differentiating both sides of Eq. (25) with respect to t yields

$$\boldsymbol{v}(\boldsymbol{x}_t, t) = \boldsymbol{u}(\boldsymbol{x}_t, t, r) + (t - r) \cdot \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}(\boldsymbol{x}_t, t, r), \tag{26}$$

where r is held fixed and $\mathrm{d}/\mathrm{d}t$ denotes the total derivative along the ODE trajectory $t \mapsto \boldsymbol{x}_t$. Motivated by the equation above, improved MeanFlow (Geng et al., 2025b) parameterizes the student average velocity field with a neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$ and defines the corrected prediction $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) := \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) + (t - r) \cdot \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r)$.

By design, matching this corrected prediction to the instantaneous velocity yields an objective analogous to the standard CFM objective:

$$\mathcal{L}_{\text{FlowMap}}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}_t, t, r} \left[\left\| \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, r) - \boldsymbol{v}(\boldsymbol{x}_t, t) \right\|_2^2 \right]. \tag{27}$$

D.2 Equivalence with original DMD updates

We show that the velocity-form distribution matching gradient used in the main text is equivalent to the original DMD update up to a time-dependent scalar weight. The difference is only a choice of parametrization: DMD is commonly written in score- or denoiser-prediction form, whereas our presentation writes the same update in velocity form.

Let the student clean sample be

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0), \qquad \boldsymbol{z} \sim p_1,$$

and let the perturbed sample be generated by the affine forward path

$$\hat{\boldsymbol{x}}_t = \alpha_t \hat{\boldsymbol{x}}_0 + \sigma_t \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim p_1, \quad t \in [0, 1].$$

We use Gaussian perturbation noise $\boldsymbol{\varepsilon}$ independent of $\boldsymbol{z}$, and hence independent of $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)$. This independence is required for the denoising identity below. Let q_t denote the marginal distribution of $\hat{\boldsymbol{x}}_t$ induced by the student, and let p_t denote the corresponding teacher marginal.

For a generic clean distribution p_0, let $\tilde{p}_t$ denote the perturbed marginal induced by

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\varepsilon}, \qquad \boldsymbol{x}_0 \sim p_0, \quad \boldsymbol{\varepsilon} \sim p_1.$$

Let

$$s_{\tilde{p}}(\boldsymbol{x}_t, t) := \nabla_{\boldsymbol{x}_t} \log \tilde{p}_t(\boldsymbol{x}_t)$$

be the score of the perturbed marginal, and let

$$\boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t) := \mathbb{E}_{\tilde{p}}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$$

be the optimal clean estimator. The standard denoising identity gives

$$s_{\tilde{p}}(\boldsymbol{x}_t, t) = \frac{\alpha_t \boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_t}{\sigma_t^2}. \tag{28}$$

Indeed,

$$\nabla_{\boldsymbol{x}_t} \tilde{p}_t(\boldsymbol{x}_t) = \int -\frac{\boldsymbol{x}_t - \alpha_t \boldsymbol{x}_0}{\sigma_t^2} q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) p_0(\boldsymbol{x}_0) \, \mathrm{d}\boldsymbol{x}_0,$$

and dividing by $\tilde{p}_t(\boldsymbol{x}_t)$ yields Eq. (28).

Now define the marginal velocity field associated with $\tilde{p}_t$ as

$$\boldsymbol{v}_{\tilde{p}}(\boldsymbol{x}_t, t) := \mathbb{E}_{\tilde{p}}[\dot{\alpha}_t \boldsymbol{x}_0 + \dot{\sigma}_t \boldsymbol{\varepsilon} \mid \boldsymbol{x}_t].$$

Since

$$\mathbb{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{x}_t] = \frac{\boldsymbol{x}_t - \alpha_t \boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t)}{\sigma_t},$$

we obtain

$$\boldsymbol{v}_{\tilde{p}}(\boldsymbol{x}_t, t) = \dot{\alpha}_t \boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t) + \dot{\sigma}_t \frac{\boldsymbol{x}_t - \alpha_t \boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t)}{\sigma_t} \tag{29}$$

$$= \frac{\dot{\sigma}_t}{\sigma_t} \boldsymbol{x}_t + \left(\dot{\alpha}_t - \frac{\alpha_t \dot{\sigma}_t}{\sigma_t} \right) \boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t). \tag{30}$$

Substituting Eq. (28), i.e.

$$\boldsymbol{\mu}_{\tilde{p}}(\boldsymbol{x}_t, t) = \frac{\boldsymbol{x}_t + \sigma_t^2 s_{\tilde{p}}(\boldsymbol{x}_t, t)}{\alpha_t},$$

gives

$$\boldsymbol{v}_{\tilde{p}}(\boldsymbol{x}_t, t) = a_t \boldsymbol{x}_t + b_t s_{\tilde{p}}(\boldsymbol{x}_t, t), \tag{31}$$

where

$$a_t = \frac{\dot{\sigma}_t}{\sigma_t} + \frac{1}{\alpha_t}\left(\dot{\alpha}_t - \frac{\alpha_t \dot{\sigma}_t}{\sigma_t}\right), \qquad b_t = \frac{\sigma_t^2}{\alpha_t}\left(\dot{\alpha}_t - \frac{\alpha_t \dot{\sigma}_t}{\sigma_t}\right).$$

The term $a_t \boldsymbol{x}_t$ is independent of the perturbed marginal. Therefore, for the fake and real marginals,

$$\boldsymbol{v}_{\text{fake}}(\boldsymbol{x}_t, t) - \boldsymbol{v}_{\text{real}}(\boldsymbol{x}_t, t) = b_t \left(s_{\text{fake}}(\boldsymbol{x}_t, t) - s_{\text{real}}(\boldsymbol{x}_t, t)\right). \tag{32}$$

Thus the score difference used by DMD and the velocity difference used in our presentation are equivalent up to the scalar b_t.

Consider the reverse-KL objective over perturbed marginals,

$$\mathcal{L}_{\text{rKL}}(\boldsymbol{\theta}) = \mathbb{E}_t \left[D_{\text{KL}}\left(q_t \,\|\, p_t\right)\right].$$

The DMD pathwise gradient has the form

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rKL}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}}\left[\left(s_{\text{fake}}(\hat{\boldsymbol{x}}_t, t) - s_{\text{real}}(\hat{\boldsymbol{x}}_t, t)\right)^\top \frac{\partial \hat{\boldsymbol{x}}_t}{\partial \boldsymbol{\theta}}\right]. \tag{33}$$

Using Eq. (32), we can rewrite this as

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rKL}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}}\left[\frac{1}{b_t}\left(\boldsymbol{v}_{\text{fake}}(\hat{\boldsymbol{x}}_t, t) - \boldsymbol{v}_{\text{real}}(\hat{\boldsymbol{x}}_t, t)\right)^\top \frac{\partial \hat{\boldsymbol{x}}_t}{\partial \boldsymbol{\theta}}\right]. \tag{34}$$

The scalar $1/b_t$ depends only on the noise schedule and can be absorbed into the time-dependent weighting function. Therefore,

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rKL}}(\boldsymbol{\theta}) \propto \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}}\left[w_t \left(\boldsymbol{v}_{\text{fake}}(\hat{\boldsymbol{x}}_t, t) - \boldsymbol{v}_{\text{real}}(\hat{\boldsymbol{x}}_t, t)\right)^\top \frac{\partial \hat{\boldsymbol{x}}_t}{\partial \boldsymbol{\theta}}\right], \tag{35}$$

which is the velocity-form gradient used in the main text.

For the linear interpolation path $\alpha_t = 1 - t$ and $\sigma_t = t$, we have

$$\boldsymbol{v}_{\tilde{p}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{\varepsilon} - \boldsymbol{x}_0 \mid \boldsymbol{x}_t],$$

and Eq. (32) becomes

$$\boldsymbol{v}_{\text{fake}}(\boldsymbol{x}_t, t) - \boldsymbol{v}_{\text{real}}(\boldsymbol{x}_t, t) = -\frac{t}{1-t}\left(s_{\text{fake}}(\boldsymbol{x}_t, t) - s_{\text{real}}(\boldsymbol{x}_t, t)\right). \tag{36}$$

Hence, under the linear schedule, the two gradients differ only by the scalar factor $(t-1)t^{-1}$, which is absorbed into w_t and depends on the sign convention for velocity definition.

Finally, we show that the fake velocity model trained in the main text is equivalent to the fake score model in DMD. Original DMD trains a fake denoiser by

$$\mathcal{L}_{\text{denoise}}(\boldsymbol{\phi}; \boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t, t) - \hat{\boldsymbol{x}}_0\right\|_2^2\right],$$

with the generator $\boldsymbol{\theta}$ frozen. The population minimizer is $\boldsymbol{\mu}^\star_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t) = \mathbb{E}[\hat{\boldsymbol{x}}_0 \mid \hat{\boldsymbol{x}}_t]$. Equivalently, if we parameterize the fake critic as a velocity field and train it with the flow-matching objective

$$\mathcal{L}_{\text{fake}}(\boldsymbol{\phi}; \boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{z},\boldsymbol{\varepsilon}}\left[\left\|\boldsymbol{v}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}_t, t) - (\dot{\alpha}_t \hat{\boldsymbol{x}}_0 + \dot{\sigma}_t \boldsymbol{\varepsilon})\right\|_2^2\right],$$

then its population minimizer is $\boldsymbol{v}^\star_{\boldsymbol{\phi}}(\boldsymbol{x}_t, t) = \mathbb{E}\left[\dot{\alpha}_t \hat{\boldsymbol{x}}_0 + \dot{\sigma}_t \boldsymbol{\varepsilon} \mid \hat{\boldsymbol{x}}_t\right] = \boldsymbol{v}_{\text{fake}}(\boldsymbol{x}_t, t)$.

According to Eq. (31), $\boldsymbol{\mu}^\star_{\boldsymbol{\phi}}$, s_{fake}, and $\boldsymbol{v}_{\text{fake}}$ are related by deterministic affine transformations at each time t. Therefore, training a fake velocity model by the flow-matching objective is equivalent to training the fake score model used in DMD, up to parametrization. Consequently, the velocity-form update in Eq. (35) optimizes the same distribution matching objective as the original DMD update.

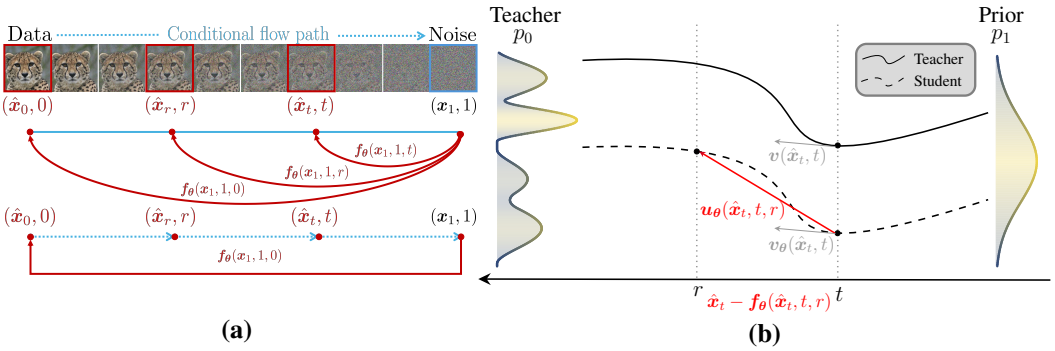


Figure 8: Overview of our method. Panel (a) illustrates two data-free constructions of $\hat{\boldsymbol{x}}_t$ shown by the red curves. In (b), we demonstrate how the student velocity matches the target teacher velocity. Note that teacher velocity is the only supervision signal.

E Implementation Details

Training. The backbone architectures follow SiT, which is based on DiT (Peebles & Xie, 2023) with adaLN-Zero (Peebles & Xie, 2023) for conditioning. To condition the student model on the additional variables, we use a two-layer SiLU-activated MLP whose hidden dimension matches the model's hidden dimension. We add new embeddings for r and w, and adopt adaLN-Zero for conditioning. The resulting conditioning signal is fed to each transformer block. No further architectural changes are necessary for stable and effective training.

All models are trained using the PyTorch framework. The detailed training hyperparameters are provided in Tab. 5.

Evaluation. For evaluation, we use the same PyTorch sampling and evaluation pipeline for all reported models to ensure a consistent comparison with baseline methods. Unless otherwise specified, we evaluate the 256×256 class-conditional ImageNet generation performance of the student model distilled from the SiT-XL/2+VA-REPA teacher. Because the model is trained with teacher guidance strengths sampled from $w \sim \mathcal{U}(1, 2)$, we sweep the guidance strength within this range at evaluation time and report the best-performing setting. The final results are reported in Tab. 1.

For the SiT-XL/2 teacher setting, we currently specify only the corresponding t and r sampling distributions, as shown in Tab. 5. The remaining settings and evaluation results are left empty until stable training outcomes are obtained.

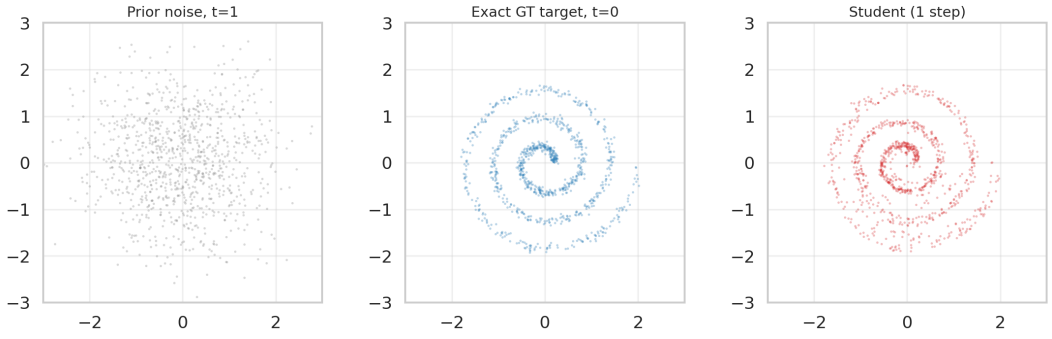


Figure 9: **Analytic 2D spiral benchmark.** Left: Gaussian prior noise at $t = 1$. Middle: exact ground-truth spiral target at $t = 0$. Right: one-step student samples after distillation.

Synthetic Dataset. We evaluate on an analytic 2D spiral benchmark, where the teacher velocity field and target distribution are available in closed form, without relying on a learned teacher checkpoint. As shown in Fig. 9, samples are transported from a Gaussian prior at $t = 1$ to a spiral target

Table 5: Training hyperparameters and model configuration. The listed distributions are the initial proposals. ‡: Teacher CFG is 1.75 for $t \in [0.125, 1]$ and 1 otherwise. The Lightning-DiT run uses post-hoc EMA for evaluation; the listed decay is the configured student EMA decay.

Task			
teacher model	SiT-XL/2+VA-REPA	SiT-XL/2	Lightning-DiT-XL/2
resolution	256×256	256×256	256×256
General			
iterations	100K	100K	100K
batch size	1024	1024	1024
optimizer	Adam	Adam	AdamW
optimizer betas	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.999)$
optimizer eps	1×10^{-8}	1×10^{-8}	1×10^{-8}
weight decay	0.0	0.0	0.0
dropout	0.0	0.0	0.0
learning rate	3×10^{-5}	3×10^{-5}	1×10^{-4}
EMA decay	0.9995	0.9995	0.999
Network			
params (M)		683	
depth		28	
DMF depth	0	0	10
hidden dim		1152	
heads		16	
patch size		2×2	
change from teacher		additional input for r and w	
Training			
t sampling	LogitNormal$(-0.4, 1.0)$	LogitNormal$(0.4, 1.0)$	LogitNormal$(-0.4, 1.6)$
r sampling	LogitNormal$(-0.4, 1.0)$	LogitNormal$(0.4, 1.0)$	LogitNormal$(1.2, 1.6)$
% of $r = t$		10%	0%
α sampling		$\mathcal{U}[0, 0.875]$	N/A
Class dropout		10%	0%
teacher CFG w		$[1, 2]$	$1.75^{\ddagger}$
m ratio		0.5	

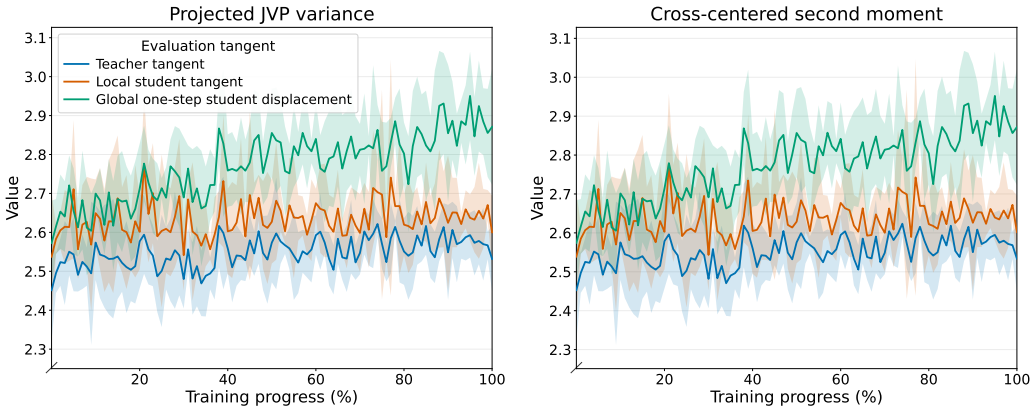
Table 6: Convergence-state evaluation results on the analytic 2D spiral benchmark.

Method	$W_2 \downarrow$	GT avg.-velocity MSE $\downarrow$	$t = 1$ endpoint MSE $\downarrow$
FreeFlow	0.190	0.317	0.0316
Ours	0.190	0.070	0.0040

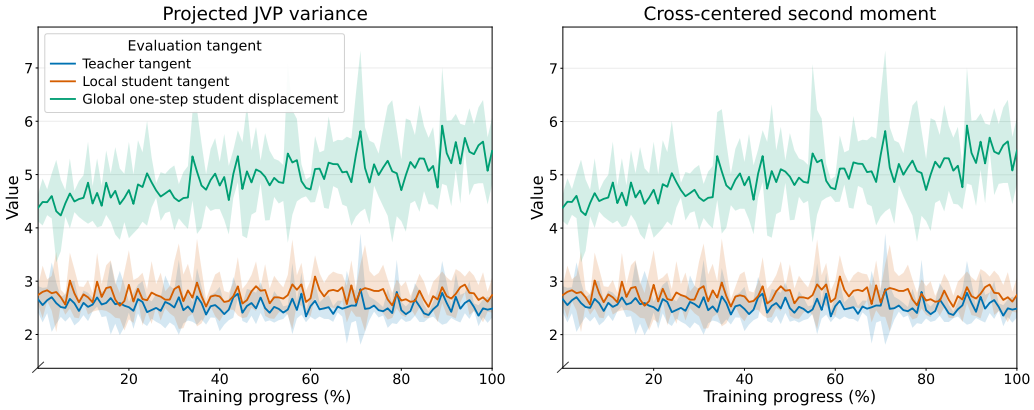
distribution at $t = 0$. This controlled low-dimensional setting allows us to evaluate not only the terminal marginal distribution but also the learned transport behavior.

Both methods use the same one-step student architecture and evaluation protocol. We compare our training objective directly against the FreeFlow objective. The converged-state results in Table 6 show that matching the terminal marginal alone is insufficient to characterize distillation quality: although both methods achieve essentially the same final W_2 score, our objective learns a substantially more accurate average velocity field and a better coupled transport map.

Figure 10 compares the projected JVP variance $\mathcal{V}_{\mathrm{JVP}}$ and the cross-centered second moment $\mathcal{M}_{\mathrm{JVP}}^{\hat{p},p}$ for the three candidate tangent directions on the 2D checkerboard (Figure 10a) and spiral (Figure 10b) datasets. Across both datasets, evaluating the JVP along the teacher tangent, $\mathrm{d}\hat{\boldsymbol{x}}_t/\mathrm{d}t = \boldsymbol{v}(\hat{\boldsymbol{x}}_t, t)$, consistently yields the smallest and most stable correction statistics throughout training. The local student tangent, $\boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, t)$, results in a modest increase in both quantities, whereas the global one-step student displacement, $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)$, exhibits substantially larger variability and an increasing trend over training. This difference is particularly pronounced on the spiral dataset, where the



(a) Results on the checkerboard dataset.



(b) Results on the spiral dataset.

Figure 10: **Effect of the JVP tangent direction on the stochasticity of the detached correction.** We compare three candidate JVP input directions on the checkerboard dataset in a and the spiral dataset in b: the teacher tangent $\boldsymbol{v}(\hat{\boldsymbol{x}}_t, t)$, the local student tangent $\boldsymbol{u}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t, t)$, and the global one-step student displacement $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}, 1, 0)$. For each dataset, the left panel reports the projected JVP variance $\mathcal{V}_{\mathrm{JVP}}$, while the right panel reports the cross-centered second moment $\mathcal{M}^{\hat{p},p}_{\mathrm{JVP}}$. Across both datasets, the teacher tangent produces the smallest and most stable statistics, whereas the global one-step displacement produces the largest variability, particularly on the spiral dataset.

global displacement remains well separated from the two state-local tangent choices. Moreover, the projected variance and cross-centered second moment exhibit nearly identical trends and rankings, suggesting that the cross-distribution contribution is relatively small in these experiments and that the observed differences are primarily driven by the intrinsic stochasticity of the JVP correction. These results support using the teacher velocity $\boldsymbol{v}(\hat{\boldsymbol{x}}_t, t)$ as the JVP input direction when constructing the detached correction target.

E.1 Text-to-image Dataset

To adapt our method on text-to-image dataset at a larger scale, some aspects need to be altered. First, we need to anchor on the teacher trajectory to prevent drifting effect. Similar to the baselines, we use 4 step during our inference. To control the noise added during training, we adopt the TDM (Luo et al., 2025) method in our data-free setting. Every time we construct the intermediate states we start with the nearest earlier teacher checkpoint and we add noise according to the forward path. We decrease the variance of our target which leads to a better final performance. In this case, we observe more diversed background and better detailed being preserved from the teacher model. With the same hardware and matching configuration, our CDMD requires approximately 7 GPU hours and 18 GB

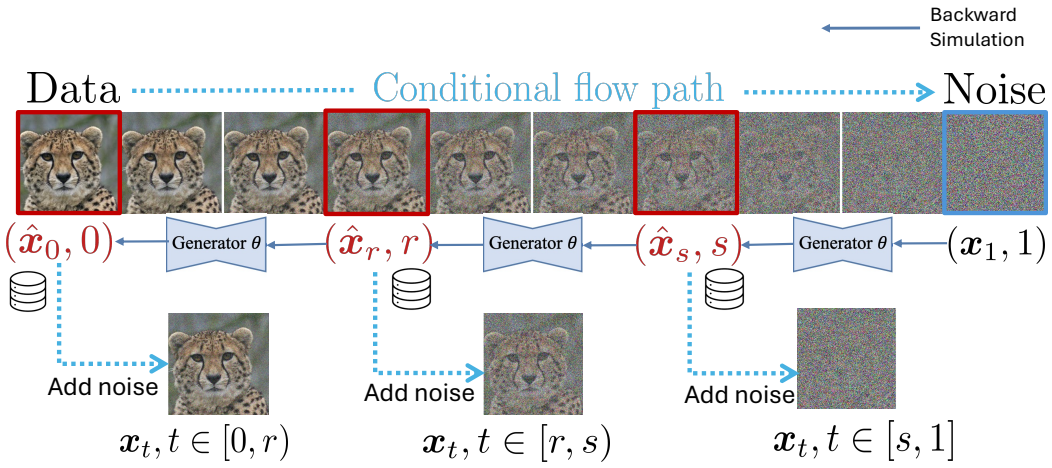


Figure 11: CDMD on Text to Image pipeline.

Algorithm 1 CDMD data-free algorithm: training.

```
t, r = sample_t_r()
m = sample_m()
e, e' = randn()

x = fn(e, 1, 0)                                  # fn(z, t, r):  student flow map
z1 = (1 - t) * x + t * e'
z2 = fn(e, 1, t)
z = m * z1 + (1 - m) * z2
z_sg = stopgrad(z)

vt = gn(z_sg, t)                                   # gn(z, t):  teacher velocity
Vt = fc(z_sg, t, r)                              # fc:  corrected velocity on fn
error = Vt - stopgrad(vt)
loss = metric(error)
```

of peak memory. Its cost is higher than CD but lower than the reported costs of Diff-Instruct, DMD2, and MFD. Note that our method performs significantly better than CD according to Table 3.

Table 7: **Training cost on SANA-1.6B.** Approximate wall-clock cost and peak memory on one NVIDIA RTX PRO 6000 GPU.

Method	GPU hours	Peak memory (GB)
Diff-Instruct (VSD) (Luo et al., 2023b)	∼22	∼20
Consistency Distillation (CD) (Song et al., 2023)	∼4	∼15
DMD2 (Yin et al., 2024a)	∼30	∼50
MFD (Zhao et al., 2026)	∼25	∼20
CDMD (ours)	∼7	∼18

F Ablation on Inference-time scaling

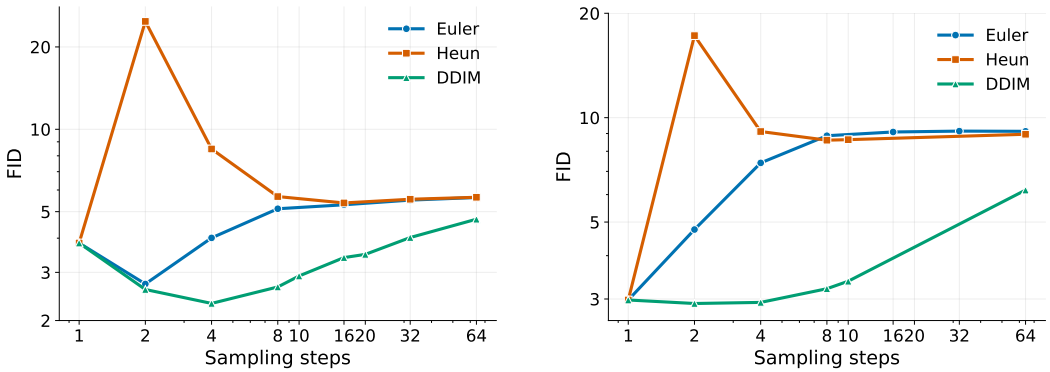
We further study inference-time scaling by comparing Euler, Heun, and DDIM samplers under different classifier-free guidance scales. Figure 12 reports FID as a function of the number of sampling steps for $w = 1.5$ and $w = 2.0$.

The ablation suggests that increasing the inference-time sampling budget does not automatically improve our distilled model. DDIM gives the strongest low-step performance in both guidance settings and remains especially stable from one to four steps. Euler is competitive at one or two steps for $w = 1.5$, but its FID increases as more steps are used and then plateaus. Heun is the least stable in the low-step regime: it exhibits a sharp FID spike at two steps and only recovers after additional

Table 8: **Reference results on class-conditional ImageNet256×256. Left**: Representative 1-NFE and multi-NFE diffusion/flow models. **Right**: Representative GANs and autoregressive/masking models. All numbers are reported with CFG when applicable, and ×2 in NFE indicates that CFG doubles the number of function evaluations at inference time.

Method	# Params	NFE	FID↓
1-NFE Diffusion / Flow from Scratch			
Shortcut-XL/2 (Frans et al., 2024)	675M	1	10.60
IMM-XL/2 (Zhou et al., 2025a)	675M	1×2	7.77
MeanFlow-XL/2 (Geng et al., 2025a)	676M	1	3.43
TVM (Zhou et al., 2025b)	678M	1	3.29
SoFlow-XL/2 (Luo et al., 2026)	675M	1	2.96
DMF-XL/2 (Lee et al., 2025)	675M	1	2.16
iMF-XL/2 (Geng et al., 2025b)	675M	1	1.72
Multi-NFE Diffusion / Flow			
ADM-G (Dhariwal & Nichol, 2021)	554M	250×2	4.59
LDM-4-G (Rombach et al., 2022)	400M	250×2	3.60
DiT-XL/2 (Peebles & Xie, 2023)	675M	250×2	2.27
SiT-XL/2 (Ma et al., 2024)	675M	250×2	2.06
SiT-XL/2+REPA (Yu et al., 2024)	675M	250×2	1.42
SiD2 (Hoogeboom et al., 2025)	–	512×2	1.38
LightningDiT (Yao et al., 2025)	675M	250×2	1.35
DDT-XL/2 (Wang et al., 2025)	675M	250×2	1.26

Method	# Params	NFE	FID↓
GANs			
BigGAN (Brock et al., 2018)	112M	1	6.95
GigaGAN (Kang et al., 2023)	569M	1	3.45
StyleGAN-XL (Sauer et al., 2022)	166M	1	2.30
Autoregressive / Masking			
MaskGIT (Chang et al., 2022)	227M	8	6.18
RQ-Transformer (Lee et al., 2022)	3.8B	256×2	3.80
STARFlow (Gu et al., 2025)	1.4B	1024×2	2.40
LlamaGen-3B (Sun et al., 2024)	3.1B	256×2	2.18
VAR-*d*30 (Tian et al., 2024)	2B	10×2	1.92
MAR-H (Li et al., 2024)	943M	256×2	1.55
RAR-XXL (Yu et al., 2025)	1.5B	256×2	1.48
Data-based Distillation			
Teacher: SiT-XL/2 (FID = 2.06)			
SDEI (Liu & Yue, 2025)	675M	8	2.46
FACM (Peng et al., 2025)	675M	2	2.07
Teacher: SiT-XL/2+REPA (FID = 1.37)			
π-Flow (Chen et al., 2025)	675M	1 2	2.85 1.97



(a) $w = 1.5$ (b) $w = 2.0$

Figure 12: Sampler comparison for inference-time scaling under classifier-free guidance scales $w = 1.5$ and $w = 2.0$. The student network is trained with SiT+VA-REPA teacher. We report FID across different numbers of sampling steps for Euler, Heun, and DDIM. DDIM gives the most stable low-step behavior, while Heun shows a sharp degradation at two steps and Euler becomes less favorable as the sampling budget increases.

steps. The effect is more pronounced under stronger guidance, where both Euler and Heun drift to noticeably worse FID than DDIM as the sampling budget grows.

These results give useful guidance for scaling inference time in our model. The interval-conditioned parameterization supports multi-step sampling, but the benefit depends strongly on the sampler and guidance scale. In this setting, DDIM is the most reliable choice for converting extra inference steps into stable sample quality, while higher-order or ODE-style updates can amplify student–teacher mismatch at small step counts. Thus, inference-time scaling should be treated as a solver-selection problem rather than simply increasing the number of steps.

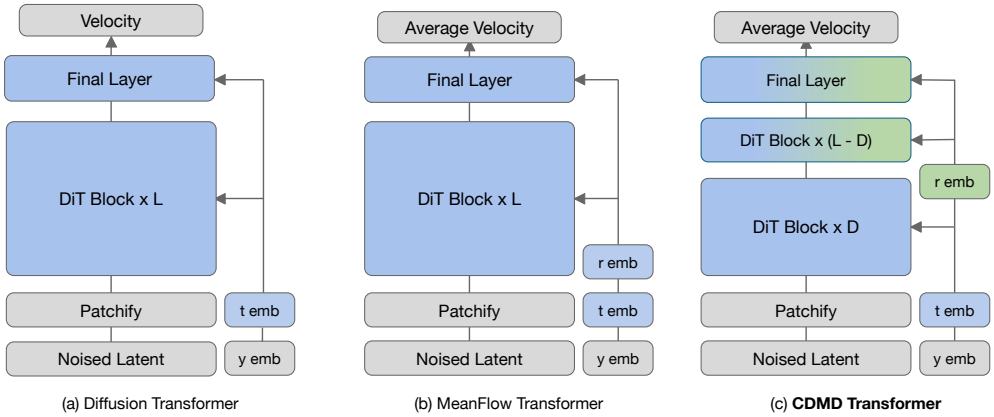


Figure 13: High level comparison of Diffusion Transformer architectures

G CDMD WITH GUIDANCE

Classifier-free guidance (CFG) (Ho & Salimans, 2022) is significant to class-conditional image generation because it facilitates conditional generation by extrapolating from unconditional to conditional predictions. Since the most pre-trained diffusion baselines are evaluated with CFG, we formulate our distillation objective to match the guided teacher dynamics rather than the unguided velocity field alone. Let $\boldsymbol{v}(\boldsymbol{x}_t, t \mid \boldsymbol{c})$ and $\boldsymbol{v}(\boldsymbol{x}_t, t \mid \varnothing)$ denote the conditional and unconditional teacher velocity fields, respectively. For a guidance scale $w \geq 1$, the guided velocity field is defined as $\boldsymbol{v}^{\text{cfg}}(\boldsymbol{x}_t, t \mid \boldsymbol{c}) = (1-w)\,\boldsymbol{v}(\boldsymbol{x}_t, t \mid \varnothing) + w\,\boldsymbol{v}(\boldsymbol{x}_t, t \mid \boldsymbol{c})$.

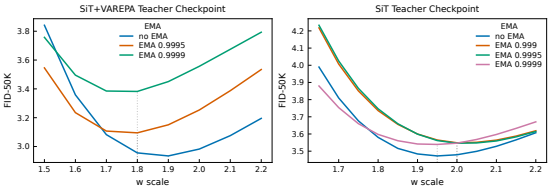


Figure 14: FID-50K versus classifier-free guidance scale w condition for students distilled from two teacher checkpoints. Left: SiT+VAREPA teacher checkpoint. Right: SiT teacher checkpoint. Different curves compare various EMA settings.

In addition, we can reinforce this signal by incorporating student guidance with scale $\alpha \geq 0$ during training to support 1-NFE sampling (Geng et al., 2025a). Compared with standard CFG sampling, which evaluates both conditional and unconditional branches at each solver step and therefore doubles the NFE count, this training-time guided target lets the distilled student absorb part of the conditional guidance signal before inference. Specifically, we define the enhanced guided velocity field as follows:

$$\boldsymbol{v}^{\text{cfg++}}(\boldsymbol{x}_t, t \mid \boldsymbol{c}) = \boldsymbol{v}^{\text{cfg}}(\boldsymbol{x}_t, t \mid \boldsymbol{c}) + \alpha\Big(\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, t \mid \boldsymbol{c}) - \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, t \mid \varnothing)\Big), \tag{37}$$

where $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, t \mid \boldsymbol{c})$ approximates the ground-truth instantaneous marginal velocity field under condition $\boldsymbol{c}$. This additional term further amplifies the conditional signal while retaining the stabilizing effect of the unconditional component. Following the flexible guidance in iMF (Geng et al., 2025b), we allow the network to condition on the guidance term.

Guidance and tangent choice. The teacher velocity enters the objective twice: as the regression target and as the tangent in the JVP. We find that using the CFG velocity $\boldsymbol{v}^{\text{cfg}}(\boldsymbol{x}_t, t \mid \boldsymbol{c})$ as the tangent and $\boldsymbol{v}^{\text{cfg++}}(\boldsymbol{x}_t, t \mid \boldsymbol{c})$ as the target performs best. Fig. 14 shows that the best w depends on both the teacher checkpoint and the EMA rate. Moderate w gives the best FID in most settings, while large w or strong EMA smoothing degrades it. We also observe that student guidance is difficult to tune with Lightning-DiT as the teacher. It sometimes produces a high IS with exaggerated class features.

H LIMITATIONS AND BROADER IMPACT

A key limitation of the current formulation is the absence of meaningful constraints on the total derivative term $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(\boldsymbol{x}_t, t, r) = \boldsymbol{v}(\boldsymbol{x}_t, t)D_{\boldsymbol{x}_t}\boldsymbol{u} + \partial_t \boldsymbol{u}$. To date, this component has not been rigorously analyzed. In prior work, it is typically detached from the computational graph and excluded from backpropagation, effectively treating it as a constant. However, empirical observations reveal that its norm is large and fluctuates substantially throughout training, suggesting that this simplification may be problematic. A plausible direction for regularization is to leverage a well-trained reference velocity field $\boldsymbol{v}(\boldsymbol{x}_t, t)$, encouraging the derivative term to exhibit similar, more stable behavior. In practice, not only is the dataset private, but the teacher model weights are also inaccessible. Hence, we cannot initialize the student model with the teacher model weights, which is a common practice in distillation work. This makes training more difficult and less stable. We hope to explore better initialization strategies in the future.

I IMAGE PROMPTS

1. **Tram:** watercolor of Istanbul rainy night with the tramway
2. **Portrait:** portrait photo of a girl, photograph, highly detailed face, depth of field
3. **Neon bear:** a bear riding a skateboard, 4k, high resolution, with a Neon sign at the back saying "CDMD"
4. **Warrior:** Artstation Assassin S Creed Iv Black Flag Concept Art Martin Deschambault
5. **Shiba Inu:** a shiba inu wearing a beret and black turtleneck
6. **Llama:** A photo of llama wearing sunglasses standing on the deck of a spaceship with the Earth in the background.
7. **Pirate ship:** Pirate ship trapped in a cosmic maelstrom nebula, rendered in cosmic beach whirlpool engine, volumetric lighting, spectacular, ambient lights, light pollution, cinematic atmosphere, art nouveau style, illustration art artwork by SenseiJaye, intricate detail, 2k
8. **Astronaut:** Astronaut in a jungle, cold color palette, muted colors, detailed, 4k
9. **Autumn river:** Fascinating autumn forest scene with a river and foliage.
10. **Capybara:** a capybara made of voxels sitting in a grass field