VA-TDM: A UNIFIED VIEW FOR IMPROVING FEW-STEP DIFFUSION MODELS BY VELOCITY ADVERSARIAL DISTILLATION

Anonymous authors
Paper under double-blind review

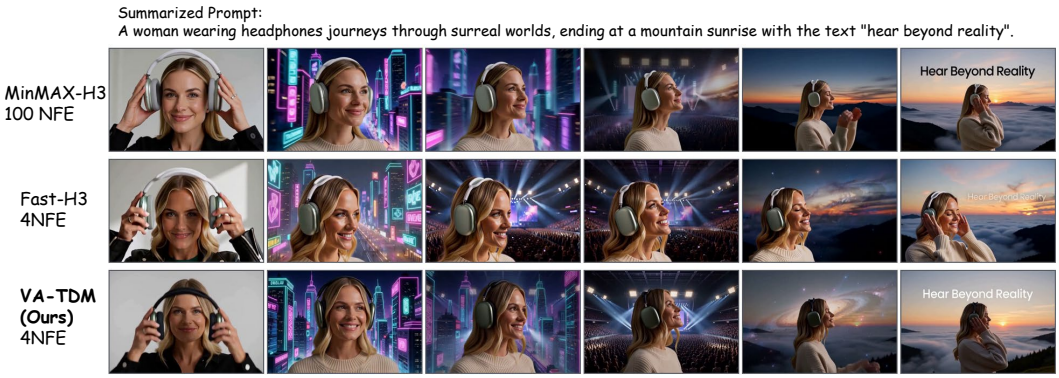

Figure 1: Comparison of VA-TDM-H3 and competing baselines. Our VA-TDM delivers better visual quality and suffers less oversaturation than the few-step competitor. More samples in Appendix E.

ABSTRACT

Few-step distillation offers efficient generation with diffusion and flow models, but preserving quality with only a few sampling steps remains challenging. Distribution-matching methods via score distillation provide a strong foundation, yet they rely on accurately tracking the student's changing distribution. At intermediate generation steps in few-step sampling, student-score estimation can be difficult, and commonly used estimators may describe a different distribution from the one used to update the student. Flow matching allows a more flexible approach to supervision through velocity fields constructed from the student's sampling process. Building on this perspective, we introduce Velocity Adversarial Distillation (VAD), which uses a learned critic to compare student and teacher fields and turn their disagreement into feedback for learning. VAD unifies score distillation, Fisher-based matching, and moment matching, also explaining existing heuristic MMD updates and delivering new training objectives. We further develop VA-TDM, which follows trajectory distribution matching: providing supervision for partial denoising along the student trajectory. Beyond this, VA-TDM combines transition-aware online velocity learning and dual-form VAD supervision across intermediate student transitions. Extensive results show that VA-TDM is powerful for learning few-step generators, achieving state-of-the-art four-step performance in both text-to-image and text-to-video generation across various backbones and scaling effectively to 33B MinMAX-H3. In particular, our VA-TDM-H3 notably outperforms the few-step Fast-H3 and is closer to the original many-step H3 according to human preference.

1 INTRODUCTION

Few-step distillation seeks to make diffusion and flow models efficient without sacrificing generation quality (Ho et al., 2020; Esser et al., 2024). A central challenge is to provide effective supervision when a small number of student transitions must replace a long teacher trajectory. Trajectory distillation learns from individual teacher trajectories (Song et al., 2023), whereas distribution matching

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.

allows the student to learn its own mapping by aligning student and teacher distributions (Luo et al., 2023; Yin et al., 2023). Trajectory Distribution Matching (TDM) (Luo et al., 2025) extends this principle to intermediate transitions. Each step then learns a partial denoising task, distributing supervision across the student trajectory and providing a natural foundation for few-step distillation.

For intermediate supervision to be effective, however, the training signal must faithfully reflect the distribution produced at each student step. In score-based methods, generator updates depend on the difference between teacher and student scores, requiring an auxiliary model to track the evolving student distribution. Intermediate transitions complicate this estimation: the student's clean prediction and retained residual can be correlated, so perturbing its intermediate output generally differs from adding fresh Gaussian noise to the clean prediction. Our analysis of TDM shows why this distinction matters. Direct score estimation for the perturbed intermediate output can involve high-variance targets, while clean-target denoising can estimate a different distribution from the one used for generator supervision.

These observations motivate supervision that explicitly accounts for the student's sampling process while allowing alternatives to its noisy-marginal score. Velocity targets offer this flexibility because they can be defined through paths with coupled endpoints (Liu et al., 2022), including dependencies inherited from a student transition. We build on this flexibility to introduce *Velocity Adversarial Distillation* (VAD), a framework for constructing student velocity fields and using their discrepancies from the teacher to guide distillation. Two prototypical velocity constructions are derived within VAD. *A-form* is a rectified velocity that captures the expected clean prediction conditioned on a perturbed sample; *B-form* is a score-equivalent Gaussian velocity for the student's perturbed distribution. The two forms thus express conditional-moment and distributional supervision within the same sampling process. A vector-valued critic estimates the discrepancy between the chosen student field and the teacher field, and the generator learns from this discrepancy.

The resulting framework provides a unified perspective along two axes: *which field to match* and *how to use its discrepancy to update the generator*. With different generator updates, A-form connects to conditional moment matching, while B-form connects reverse-KL score distillation and Fisher-based field matching. These connections also clarify the relationship between an intended objective and its practical update. In particular, VAD explains the heuristic update of MMD (Salimans et al., 2024) and derives an alternative surrogate that recovers the exact gradient of MMD's desired objective.

To use both types of supervision along a few-step trajectory, we develop *VA-TDM*, retaining TDM's supervision of intermediate transitions. VA-TDM combines A-form and B-form through a shared auxiliary backbone and blends complementary generator updates. Because samples from different student steps can follow different distributions even at the same noise level, we also condition field estimation on the transition being supervised. This provides different and accurate supervision signals among different steps.

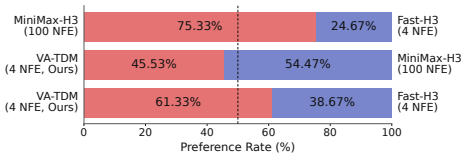

Figure 2: Human preference on video synthesis.

Overall, our unified VAD framework connects existing distillation updates and derives new objectives, which VA-TDM combines with complementary velocity fields and transition-aware estimation for four-step generation, achieving state-of-the-art performance across backbones and tasks. On SD3.5-Medium, VA-TDM outperforms the compared few-step baselines across prompt following, image quality, and preference metrics, improving GenEval from TDM's 0.61 to 0.67 while increasing sample diversity. On FLUX.2-Klein-Base 4B, it surpasses the released four-step model, raising GenEval from 0.78 to 0.80 and OCR similarity from 0.64 to 0.69, with diversity measured by LPIPS improving from 0.47 to 0.56. Most notably, VA-TDM scales to MiniMax-H3 (MiniMax, 2026), a state-of-the-art open-source audio-video model with 33B parameters. In our user study (Fig. 2), VA-TDM wins 61.33% of comparisons against Fast-H3 at the same 4-NFE budget and achieves a 45.53% preference rate against the 100-NFE teacher, compared with 24.67% for Fast-H3. These results demonstrate that VA-TDM substantially narrows the perceptual quality gap to the teacher while requiring only 4-NFE.

2 Preliminaries

Diffusion and Flow Models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) use the Gaussian forward process $\mathbf{x}_\tau = \alpha_\tau \mathbf{x}_0 + \sigma_\tau \boldsymbol{\varepsilon}$, where $\mathbf{x}_0 \sim q_0$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independently,

and q_τ denotes the perturbed marginal. Time $\tau \in [0,1]$ runs from clean data to the noise prior; conditioning on $\mathbf{c}$ is omitted. For $\sigma_\tau > 0$, denoising score matching (DSM) learns the marginal score by regressing onto the conditional target:

$$\mathbf{s}^q_\tau(\mathbf{x}_\tau) \coloneqq \nabla_{\mathbf{x}_\tau} \log q_\tau(\mathbf{x}_\tau) = \mathbb{E}_{q_0(\mathbf{x}_0|\mathbf{x}_\tau=\mathbf{x})}[\nabla_{\mathbf{x}_\tau} \log q_\tau(\mathbf{x}|\mathbf{x}_0)]\,. \tag{1}$$

Rectified flow (Liu et al., 2022) uses the linear schedule $\alpha_\tau = 1-\tau$, $\sigma_\tau = \tau$ and velocity $\mathbf{u}_\tau(\mathbf{x}) = \mathbb{E}[\boldsymbol{\varepsilon} - \mathbf{x}_0 \mid \mathbf{x}_\tau = \mathbf{x}]$. For Gaussian forward processes, it is equivalent to diffusion models.

Trajectory Distribution Matching (TDM). TDM (Luo et al., 2025) matches intermediate student and teacher distributions along a K-step student trajectory $0 = t_0 < \cdots < t_K = 1$. For a transition $t = t_k \to s = t_{k-1}$, perturbing $\mathbf{x}_s \sim p_{\theta,k}(\mathbf{x}_s)$ with $q(\mathbf{x}_\tau \mid \mathbf{x}_s)$, gives student marginal $p^{\text{path}}_{\theta,\tau}$. Learning is performed by reverse-KL divergence (Luo et al., 2023) along the student trajectory:

$$\nabla_\theta \mathcal{L}_{\text{TDM}} = \mathbb{E}_{k,t>\tau>s}\mathbb{E}_{q(\mathbf{x}_\tau|\mathbf{x}_s)p_{\theta,k}(\mathbf{x}_s)}\left[w_{k,\tau}\left(\frac{\partial \mathbf{x}_\tau}{\partial \theta}\right)^\top \left(\nabla_{\mathbf{x}_\tau}\log p^{\text{path}}_{\theta,\tau}(\mathbf{x}_\tau) - \mathbf{s}^q_\tau(\mathbf{x}_\tau)\right)\right]. \tag{2}$$

Teacher scores are supplied by the pretrained model; student scores require online estimation.

3 METHOD

Problem Setup. We consider distilling a pretrained conditional diffusion or flow model into an efficient few-step generator. Let $q_t(\mathbf{x}_t \mid \mathbf{c})$ and $\mathbf{u}_t(\mathbf{x}_t, \mathbf{c})$ denote the teacher's marginal distribution and velocity field at time $t \in [0,1]$, respectively, where $t = 0$ corresponds to clean samples and $t = 1$ to the noise prior. Given a condition $\mathbf{c}$ and a desired number of sampling steps K, our goal is to learn a student that transforms noise into a sample through K transitions while matching the teacher's marginal distributions along the generation trajectory.

Overview. We introduce *Velocity Adversarial Distillation* (VAD) as a unified framework for distilling diffusion models through aligning student and teacher velocity fields by an adversarial game. Sec. 3.1 motivates this formulation by examining the challenges of student score estimation and introduces two constructions of student velocity. Sec. 3.2 develops the adversarial velocity game, characterizes its matching equilibrium, and establishes connections to existing distillation methods. Sec. 3.3 presents practical generator and critic updates and applies the framework along the student trajectory, with transition-aware field estimation.

3.1 FROM SCORE DISTILLATION TO VELOCITY DISTILLATION

We focus on learning a one-step marginal transition from $\mathbf{x}_t$ to a less noisy $\mathbf{x}_s$, where $0 \le s < t \le 1$, whose output distribution $p_{\theta,s}$ should match the teacher marginal q_s. Matching an intermediate distribution requires only partial denoising at each step, easing the learning task compared with requiring a clean output at every step, and is more efficient and effective (Luo et al., 2025).

Under the rectified-flow schedule, the student predicts a clean anchor and its residual given $\mathbf{x}_t$:

$$\hat{\mathbf{x}}_0 = g_\theta(\mathbf{x}_t, t) = \mathbf{x}_t - t v_\theta(\mathbf{x}_t, t), \qquad \hat{\boldsymbol{\varepsilon}} = \frac{\mathbf{x}_t - (1-t)\hat{\mathbf{x}}_0}{t}. \tag{3}$$

Given $\hat{\mathbf{x}}_0$ and $\hat{\boldsymbol{\varepsilon}}$, we can parameterize a DDIM-style intermediate transition (Song et al., 2021):

$$\mathbf{x}_s = (1-s)\hat{\mathbf{x}}_0 + \eta s\,\hat{\boldsymbol{\varepsilon}} + s\sqrt{1-\eta^2}\mathbf{z}_s \sim p_\theta(\mathbf{x}_s|\mathbf{x}_t) \triangleq \mathcal{N}((1-s)\hat{\mathbf{x}}_0 + \eta s\,\hat{\boldsymbol{\varepsilon}}, s^2(1-\eta^2)\mathbf{I}), \tag{4}$$

where $0 \le \eta \le 1$ and $\mathbf{z}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that the clean anchor is a parameterization of the transition; the desired matching objective is imposed on $\mathbf{x}_s$, not separately on $\hat{\mathbf{x}}_0$. By default, we let $\eta = 1$, which forms a deterministic transition, while its outputs remain random across inputs.

Distributional Supervision at a Perturbed Time. To supervise the intermediate distribution, existing works in the line of score distillation (Luo et al., 2023; 2025) further perturb $\mathbf{x}_s$ to $\mathbf{x}_\tau$, where $\tau > s$. This smooths the distributions and makes the teacher's knowledge at additional noise levels available for supervision. For $s < \tau < 1$, we use the teacher-compatible kernel

$$q(\mathbf{x}_\tau \mid \mathbf{x}_s) = \mathcal{N}(\mathbf{x}_\tau; \alpha_{\tau|s}\mathbf{x}_s, \beta^2_{\tau|s}\mathbf{I}), \quad \alpha_{\tau|s} = \frac{1-\tau}{1-s}, \qquad \beta^2_{\tau|s} = \tau^2 - \alpha^2_{\tau|s}s^2. \tag{5}$$

Thus $\mathbf{x}_\tau = \alpha_{\tau|s}\mathbf{x}_s + \beta_{\tau|s}\mathbf{z}_\tau$, where $\mathbf{z}_\tau \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The full sampling chain factors into $q(\mathbf{x}_s \mid \hat{\mathbf{x}}_0, \hat{\boldsymbol{\varepsilon}}, \eta)$ and $q(\mathbf{x}_\tau \mid \mathbf{x}_s)$. Unless otherwise specified, expectations are taken over sampled τ, this student transition and subsequent perturbation, with conditioning fixed. Let $p_{\theta,s}$ denote

the distribution of the student's intermediate output $\mathbf{x}_s$. After applying the perturbation kernel, the resulting perturbed marginal distribution is

$$p^{\text{path}}_{\theta,\tau}(\mathbf{x}_\tau) = \int q(\mathbf{x}_\tau \mid \mathbf{x}_s)\, p_{\theta,s}(\mathbf{x}_s)\, \mathrm{d}\mathbf{x}_s. \tag{6}$$

This is the student distribution whose score is required for reverse-KL supervision at time τ. Because $\mathbf{x}_s$ retains the residual inherited from the student transition, this distribution generally differs from one obtained by adding independent Gaussian noise directly to $\hat{\mathbf{x}}_0$.

If the optimization is formed by reverse-KL following previous works (Luo et al., 2025), we would require the score difference $\nabla \log p^{\text{path}}_{\theta,\tau} - \nabla \log q_\tau$. The teacher supplies the latter score, whereas the former must be learned online for the current student's actual perturbation distribution.

The Difficulty of Learning the Student Score. DSM can estimate this score $\nabla \log p^{\text{path}}_{\theta,\tau}$ by regressing onto $\nabla_{\mathbf{x}_\tau} \log q(\mathbf{x}_\tau \mid \mathbf{x}_s)$. *Although population-correct, under the reference Gaussian forward process this target has larger conditional variance than standard DSM*, which regresses $\nabla_{\mathbf{x}_\tau} \log q(\mathbf{x}_\tau \mid \mathbf{x}_0)$ (Xue et al., 2025), thereby making the online score harder to learn.

Remark 1 (Mismatch in TDM)**.** *To avoid this difficulty, TDM learns its online score through clean-target denoising with importance sampling (Luo et al., 2025). Although importance sampling allows training samples to be drawn from $p^{\text{path}}_{\theta,\tau}$, the objective still estimates the score of the distribution obtained by independently perturbing $\hat{\mathbf{x}}_0$ to τ. This generally differs from the score of $p^{\text{path}}_{\theta,\tau}$ required by the generator update, introducing a distribution mismatch (See detailed derivation in Appendix B.2).*

Velocity as a Choice of Path and Coupling. This distinction motivates a shift from score distillation to velocity distillation, which can define student fields directly from the actual sampling chain. In particular, ***Velocity targets can be induced by differentiable paths with correlated or non-Gaussian endpoint couplings***, therefore giving us greater flexibility. In the student transition $t \to s$, $\hat{\mathbf{x}}_0$ and $\hat{\boldsymbol{\varepsilon}}$ share the same input and can be dependent, alongside optional fresh noise; the forward perturbation $s \to \tau$ adds independent Gaussian noise. This gives two choices: use the clean-to-perturbed coupling induced by the composite process $t \to s \to \tau$, or follow the $s \to \tau$ path alone. Throughout the remaining sections, $p_{\theta,\tau}$ abbreviates $p^{\text{path}}_{\theta,\tau}$, the perturbed marginal.

A-form: Rectified Velocity. The composite process $t \to s \to \tau$ couples the clean prediction $\hat{\mathbf{x}}_0$ to $\mathbf{x}_\tau$. Expanding Eq. (4) yields

$$\begin{aligned} \mathbf{x}_\tau &= (1-\tau)\hat{\mathbf{x}}_0 + \boldsymbol{\xi}_{\theta,s,\tau}, \\ \boldsymbol{\xi}_{\theta,s,\tau} &= \alpha_{\tau|s}\eta s\, \hat{\boldsymbol{\varepsilon}} + \alpha_{\tau|s} s\sqrt{1-\eta^2}\mathbf{z}_s + \beta_{\tau|s}\mathbf{z}_\tau. \end{aligned} \tag{7}$$

The effective noise $\boldsymbol{\varepsilon}_{\text{eff}} = \boldsymbol{\xi}_{\theta,s,\tau}/\tau$ contains the residual inherited from the student and standard Gaussian noise. This can deliver a corresponding rectified velocity coupled with the endpoint:

$$\mathbf{v}^A_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0) = \boldsymbol{\varepsilon}_{\text{eff}} - \hat{\mathbf{x}}_0 = \frac{\mathbf{x}_\tau - \hat{\mathbf{x}}_0}{\tau}. \tag{8}$$

The A-form rectified velocity uses the clean prediction as its anchor through the composite coupling. It produces reduced training variance when $s > 0$, which benefits learning of the online field. This marginal velocity is an affine reparameterization of a moment $\mathbb{E}[\mathbf{x}_0|\mathbf{x}_\tau]$.

B-form: Gaussian Velocity. The perturbation $s \to \tau$ defines a path starting at $\mathbf{x}_s$. Differentiating this path with respect to τ, and using $\mathbf{z}_\tau = (\mathbf{x}_\tau - \alpha_{\tau|s}\mathbf{x}_s)/\beta_{\tau|s}$, gives the conditional velocity

$$\mathbf{v}^B_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0) = \dot{\alpha}_{\tau|s}\mathbf{x}_s + \frac{\dot{\beta}_{\tau|s}}{\beta_{\tau|s}}\big(\mathbf{x}_\tau - \alpha_{\tau|s}\mathbf{x}_s\big). \tag{9}$$

Dots denote τ derivatives with s fixed. Its conditional average is the marginal velocity of the $s \to \tau$ perturbation path. This marginal velocity is an affine reparameterization of the score of the actual perturbed distribution (see proof in Appendix A.4).

The Induced Student Field. We use the common notation $\mathbf{v}^F_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)$ for these two conditional targets, with $F \in \{A, B\}$. The field to be compared with the teacher is

$$\bar{\mathbf{v}}^F_{\theta,\tau}(\mathbf{x}) = \mathbb{E}_{p_\theta(\mathbf{x}_s, \hat{\mathbf{x}}_0 \mid \mathbf{x}_\tau = \mathbf{x})}\left[\mathbf{v}^F_\tau(\mathbf{x} \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)\right]. \tag{10}$$

Both formulations can be computed via student samples, but their marginal fields require averaging over the implicit student distribution.

3.2 VELOCITY ADVERSARIAL DISTILLATION

Diff-Instruct has discussed that reverse-KL-based score distillation is related to GANs (Luo et al., 2023; Goodfellow et al., 2014). We further find that, after shifting from score distillation to velocity distillation, explicitly defining an adversarial game that aims to align the teacher velocity with the student velocity forms a unified framework and view, which deepens the understanding of existing methods and might inspire the design of new distillation losses.

Matching Implicit Velocity. We propose to align the student and teacher velocity fields through an adversarial game defined by their inner products with a vector-valued critic $\mathbf{f}$, i.e., $\mathbf{f}^\top \mathbf{v}$. In particular, the critic maximizes its alignment with the teacher velocity and minimizes its alignment with the student velocity, subject to a quadratic penalty. The student opposes the critic by maximizing its own velocity's alignment with the critic, with the critic held fixed.

For brevity, we write $\mathbf{v}_\tau^F \equiv \mathbf{v}_\tau^F(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)$ for the sampled conditional velocity, and let $\mathbb{E}$ denote expectation over τ and the full sampling chain $p_\theta(\mathbf{x}_t, \mathbf{x}_s, \hat{\mathbf{x}}_0) q(\mathbf{x}_\tau \mid \mathbf{x}_s)$. Since $\mathbf{f}$ depends only on $\mathbf{x}_\tau$, replacing the sampled velocity $\mathbf{v}_\tau^F$ by its conditional-mean velocity $\bar{\mathbf{v}}_\tau^F$ leaves the expected inner product unchanged:

$$\mathbb{E}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \mathbf{v}_\tau^F\big] = \mathbb{E}_{\mathbf{x}_\tau \sim p_{\theta,\tau}}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}_\tau)\big]. \tag{11}$$

This identity motivates two generator losses with the same expected value but different gradient conventions. The *pathwise update* differentiates through the full sampling process, changing both the outer sampling distribution and the student velocity at those locations. The *conditional-mean update* stops the gradient of the sampling distribution and adjusts only the student velocity field.

In particular, the adversarial game in the velocity space can be formulated as:

$$\max_{\mathbf{f}} \quad V_F(\mathbf{f}, \theta) = \mathbb{E}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \big(\mathbf{u}_\tau(\mathbf{x}_\tau) - \bar{\mathbf{v}}_{\theta,\tau}^F\big) - \tfrac{1}{2}\|\mathbf{f}(\mathbf{x}_\tau)\|^2\big], \tag{12}$$

$$\min_{\theta} \quad \mathcal{L}_{\text{path}}^F(\theta; \mathbf{f}) = -\mathbb{E}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \mathbf{v}_\tau^F\big],$$

$$\text{or} \qquad \min_{\theta} \quad \mathcal{G}_{\text{mean}}^F(\theta; \mathbf{f}) = -\mathbb{E}_{\mathbf{x}_\tau \sim p_{\text{sg}(\theta),\tau}}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}_\tau)\big]. \tag{13}$$

Here sg denotes stop-gradient. Only the outer distribution is stopped in $\mathcal{G}_{\text{mean}}^F$; $\bar{\mathbf{v}}_{\theta,\tau}^F$ retains its parameter dependence. The following lemma reveals the formulation of the optimal critic:

Lemma 2 (Optimal critic)**.** *For a fixed student θ, under the regularity and critic assumptions detailed in Appendix A, the maximizing critic is unique up to $p_{\theta,\tau}$-null sets and satisfies*

$$\mathbf{f}_F^\star(\mathbf{x}) = \mathbf{u}_\tau(\mathbf{x}) - \bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}). \tag{14}$$

The proof is given in Appendix A.1.

Matching Equilibrium. The optimal critic in Lemma 2 is the velocity residual. Its zero identifies the following equilibrium of the two-player game.

Proposition 3 (Matching equilibrium)**.** *Under the regularity, critic, and realizability assumptions in Appendix A, the game in Eq.* (12) *admits a Nash equilibrium satisfying*

$$\bar{\mathbf{v}}_{\theta^\star,\tau}^F = \mathbf{u}_\tau \quad p_{\theta^\star,\tau}\text{-almost everywhere}, \qquad \mathbf{f}^\star \equiv \mathbf{0}. \tag{15}$$

It attains zero velocity discrepancy, the minimum of the maximized critic payoff.

The proof is in Appendix A.2.

Directly computing the conditional-mean update is intractable because it requires differentiating the implicit field $\bar{\mathbf{v}}_{\theta,\tau}^F$ with respect to θ. The following lemma gives a general flow-product update through a tractable surrogate whose gradient agrees with the conditional-mean update.

Lemma 4 (General Flow-Product Update)**.** *Fix a reference parameter $\theta_0 = \bar{\theta}$ and let $\mathbf{h}_0^F(\mathbf{x}) = \bar{\mathbf{v}}_{\theta_0,\tau}^F(\mathbf{x})$. Hold the parameters of $\mathbf{f}$ and $\mathbf{h}_0^F$ fixed, while retaining their input derivatives. Define*

$$\mathcal{G}_{\text{fp}}^F(\theta; \mathbf{f}, \mathbf{h}_0) := -\mathbb{E}_{p_\theta(\mathbf{x}_t, \mathbf{x}_s, \hat{\mathbf{x}}_0) q(\mathbf{x}_\tau \mid \mathbf{x}_s)}\big[\mathbf{f}(\mathbf{x}_\tau)^\top \big(\mathbf{v}_\tau^F(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0) - \mathbf{h}_0^F(\mathbf{x}_\tau)\big)\big]. \tag{16}$$

Under the regularity conditions detailed in Appendix A.3,

$$\nabla_\theta \mathcal{G}_{\text{fp}}^F\big|_{\theta_0} = \nabla_\theta \mathcal{G}_{\text{mean}}^F(\theta; \theta_0, \mathbf{f})\big|_{\theta_0} = -\mathbb{E}_{\mathbf{x}_\tau \sim p_{\theta_0,\tau}}\Big[\mathbf{f}(\mathbf{x}_\tau)^\top \, \partial_\theta \bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}_\tau)\big|_{\theta_0}\Big]. \tag{17}$$

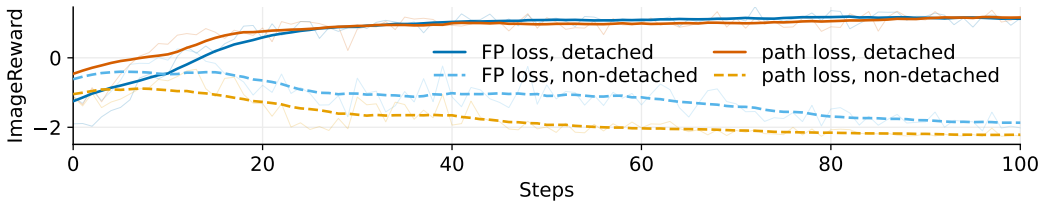

Figure 3: Comparison between detached critic and non-detached critic. We found that the non-detached critic leads to unstable training and collapse.

The proof is given in Appendix A.3.

Theorem 5 (Matching meanings and algorithmic unification)**.** *Use the compatible Gaussian kernel in Eq.* (5) *and an exact teacher. For the gradient identities, fix the input law and times* $0 \le s < t \le 1$*,* $s < \tau < 1$*, take* $\eta = 1$*, and assume exact critics and reference fields with the differentiability conditions of Appendix A.3.*

(i) Matching. *A-form velocity matching is equivalent to the conditional-moment matching*

$$\mathbf{m}_\theta(\mathbf{x}) \coloneqq \mathbb{E}[\hat{\mathbf{x}}_0 \mid \mathbf{x}_\tau = \mathbf{x}] = \mathbf{m}_q(\mathbf{x}) \coloneqq \mathbf{x} - \tau \mathbf{u}_\tau(\mathbf{x}). \tag{18}$$

B-form velocity matching is equivalent to distribution matching $p_{\theta,\tau} = q_\tau$*, and hence* $p_{\theta,s} = q_s$*.*

(ii) Equivalence at $s = 0$. *At $s = 0$, A and B coincide, with identical matching conditions and population gradients under the same generator update.*

(iii) Unifying Algorithms. *The two forms and two gradient choices yield the correspondences below, up to positive time weights.*

Form	*Pathwise w/ detached critic* $\mathrm{sg}(\mathbf{f}(\mathbf{x}_\tau))$	*Flow-product w/ non-detach critic* $\mathbf{f}(\mathbf{x}_\tau)$
A	*MMD practical update (Salimans et al., 2024)*	*Exact Moment Matching*
B	*Reverse-KL (Luo et al., 2023; 2025; Yin et al., 2023)*	*Fisher Divergence (Luo et al., 2024; Zhou et al., 2024b)*

The proof and time coefficients are in Appendix A.4.

MMD: Objective and Approximate Update. MMD (Salimans et al., 2024) aims to minimize a squared conditional-moment discrepancy to align the teacher moment and student moment. Because differentiating the student's implicit moment is intractable, it adopts a heuristic approximate gradient to update the student generator. ***Our unified framework provides both an exact gradient and a mechanistic explanation of this heuristic***: A-form flow-product updates recover the intended objective's exact gradient under MMD's fixed-outer-measure convention, while A-form pathwise updates with a detached critic recover the surrogate update as a generator update in the adversarial velocity game, up to positive time weighting on the same sampling chain. Appendix A.4 explicitly gives the target objective, the approximation leading to the surrogate, and both gradient identities.

3.3 VA-TDM: Distillation Along the Student Trajectory

Generator Learning. We observe unstable training when gradients are propagated through the critic output (Fig. 3). We therefore stop the entire critic output in the practical update. Let sg denote this stop-gradient operation and define

$$\begin{aligned} \mathcal{L}^F_{\mathrm{path,sg}} &= -\mathbb{E}_{p_\theta(\mathbf{x}_t,\mathbf{x}_s,\hat{\mathbf{x}}_0)q(\mathbf{x}_\tau|\mathbf{x}_s)}\left[\mathrm{sg}(\mathbf{f}^F(\mathbf{x}_\tau))^\top \mathbf{v}^F_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)\right], \\ \mathcal{L}^F_{\mathrm{fp,sg}} &= -\mathbb{E}_{p_\theta(\mathbf{x}_t,\mathbf{x}_s,\hat{\mathbf{x}}_0)q(\mathbf{x}_\tau|\mathbf{x}_s)}\left[\mathrm{sg}(\mathbf{f}^F(\mathbf{x}_\tau))^\top \left(\mathbf{v}^F_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0) - \bar{\mathbf{v}}^F_{\theta,\tau}\right)\right]. \end{aligned} \tag{19}$$

Note that $L^F_{\mathrm{fp,sg}}$ and $L^F_{\mathrm{path,sg}}$ correspond to different adversarial flexibility for the generator: the latter allows the generator to cheat the critic by altering both the outer sampling distribution and the velocity. Since the balance of the adversarial game is crucial for performance, we propose optimizing the hybrid loss to provide an explicit knob for controlling the balance.

$$\mathcal{L}^F_{\mathrm{mix}} = (1-\lambda)\mathcal{L}^F_{\mathrm{path,sg}} + \lambda \mathcal{L}^F_{\mathrm{fp,sg}}. \tag{20}$$

Critic Learning. With the help of Lemma 2, we parameterize the critic as $\mathbf{f}^F_\psi = \mathbf{u}_\tau - \mathbf{h}^F_\psi$ and fit the student field with

$$\mathcal{L}^F_{\mathrm{field}}(\psi) = \mathbb{E}_{p_\theta(\mathbf{x}_s,\hat{\mathbf{x}}_0)q(\mathbf{x}_\tau|\mathbf{x}_s)}\left[\|\mathbf{h}^F_\psi(\mathbf{x}_\tau) - \mathbf{v}^F_\tau(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)\|^2\right]. \tag{21}$$

Table 1: **Systematic evaluation.** Best / second-best few-step results per backbone are **bold** / underlined.

Model	NFE	Verifiable		ImgQ	Preference				Div.
		GenEval	OCR	Aest.	ImgRwd	PickScore	HPSv2	HPSv3	LPIPS
SD3.5-Medium (teacher, CFG=4.5)	50×2	0.65	0.58	5.37	0.86	0.8628	0.2805	9.78	0.61
SD3.5-Medium (teacher, CFG=7)	50×2	0.67	0.62	5.37	0.96	0.8640	0.2854	9.87	0.60
DMD2	4	0.60	0.53	5.55	1.00	0.8648	0.2977	11.18	0.53
D-DMD	4	0.55	0.56	5.51	0.99	0.8567	0.2897	10.93	0.57
DPDMD	4	0.62	0.33	5.28	0.80	0.8530	0.2707	8.13	**0.59**
TDM	4	0.61	0.57	5.58	1.01	0.8679	0.2969	11.54	0.57
VA-TDM (Ours)	4	**0.67**	**0.59**	**5.65**	**1.02**	**0.8702**	**0.3010**	**11.89**	0.58
FLUX.2-Klein-Base 4B (teacher, CFG=4.0)	50×2	0.77	0.60	5.31	0.85	0.8605	0.2768	10.28	0.56
FLUX.2-Klein-Base 4B (teacher, CFG=7)	50×2	0.80	0.64	5.38	0.93	0.8620	0.2840	10.58	0.54
FLUX.2-Klein 4B	4	0.78	0.64	5.28	1.04	**0.8746**	0.2886	11.27	0.47
VA-TDM (Ours)	4	**0.80**	**0.69**	**5.43**	**1.07**	0.8694	**0.2930**	**11.58**	**0.56**

To retain the pre-trained knowledge, we add the regularization term $\lambda_r \mathbb{E}[\|\mathbf{h}^F_\psi(\mathbf{x}_\tau) - \mathbf{u}_\tau(\mathbf{x}_\tau)\|_2^2]$, with $\lambda_r > 0$, under the same sampling distribution. The optimal minimizer then becomes

$$\mathbf{h}^F_{\psi^\star}(\mathbf{x}) = \frac{\bar{\mathbf{v}}^F_{\theta,\tau}(\mathbf{x}) + \lambda_r \mathbf{u}_\tau(\mathbf{x})}{1+\lambda_r}.$$

We therefore should correctly parameterize the critic as $\mathbf{f}^F_\psi = (1+\lambda_r)(\mathbf{u}_\tau - \mathbf{h}^F_\psi)$ to perform best.

Dual-Form Velocity Game with a Shared Backbone. The A-form produces low-variance online field learning, but it only captures the conditional-mean moment of the distribution; the B-form induces high-variance online field learning, yet it encodes the full distributional information. Since no single velocity construction enjoys both benefits, we propose to optimize the adversarial games induced by the two velocity constructions simultaneously, the generator loss becomes:

$$\mathcal{L}_{\text{mix}} = \lambda_m \mathcal{L}^A_{\text{mix}} + (1-\lambda_m)\mathcal{L}^B_{\text{mix}}, \tag{22}$$

where each term follows Eq. (20) with its corresponding velocity target. To avoid the cost of maintaining two online fields, we propose a shared-backbone strategy: the two field networks share the backbone with separate output heads. Tab. 2 shows that this sharing strategy not only reduces memory and training overhead, but also improves performance.

Distillation Along the Student Trajectory. Matching only the final output places the entire distributional correction at the end of generation. We have discussed the algorithm for learning one-step marginal transition from $\mathbf{x}_t$ to $\mathbf{x}_s$. Following TDM (Luo et al., 2025), we instead apply VAD to intermediate student transitions along the student trajectory. Let $0 = t_0 < t_1 < \cdots < t_K = 1$ be a sampling grid. Starting from $\mathbf{x}_{t_K} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the student repeatedly applies the student transition in Eq. (4) with $(t, s) = (t_k, t_{k-1})$ and $\eta = 1$. Each transition learns a distributional mapping from its current noisy state to the next state; no pointwise teacher ODE solution is needed.

Transition-Aware Field Estimation. In both our framework and existing score distillation work, it is crucial for the auxiliary score to accurately track the student distribution in order to achieve strong performance. However,*we identify an aspect that has been consistently overlooked in prior work (Yin et al., 2024; Liu et al., 2026)*: in the few-step generation setting, student transitions with different step indices generally induce distinct noisy distributions, yet existing methods typically model them with a shared auxiliary score, which theoretically precludes the possibility of precisely tracking the student distribution. TDM tries to avoid this by assigning student transitions with different step indices to disjoint perturbation intervals, which is partially effective but sacrifices part of supervision signals. To resolve this issue, ***we propose that the field network should additionally receive the identity of the transition***—for instance, in our setting, the target timestep of the student transition can be injected as extra input, i.e., learn $h_\psi(\mathbf{x}_\tau, \tau, s)$ as online velocity. Results in Tab. 2 show that this modification is essential for providing accurate supervision signals, and substantially improves diversity and fine-grained rendering of images, such as text rendering.

4 Experiments

Models and Baselines. We distill SD3.5-Medium (Esser et al., 2024), FLUX.2-Klein-Base 4B (Black Forest Labs, 2026) and MinMAX-H3 (MiniMax, 2026) into four-step generators. For



Figure 4: **Qualitative comparison** among VA-TDM and competing baselines.

SD3.5-Medium, the baselines include DMD2 (Yin et al., 2024), D-DMD (Liu et al., 2026), DPDMD (Wu et al., 2026), and TDM (Luo et al., 2025). For FLUX.2 4B, we compare with the officially released four-step model.

Evaluation Protocol. We evaluate images along four complementary dimensions. *Prompt following* is assessed using GenEval (Ghosh et al., 2023) for compositional generation and OCR for visual text rendering. *Image quality* is measured by Aesthetic Score (Schuhmann et al., 2022), while *preference* is assessed with ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2, and HPSv3; these quality and preference metrics are evaluated on DrawBench (Saharia et al., 2022). Finally, *diversity* is measured by average pairwise LPIPS (Zhang et al., 2018) between independently generated images by the same prompt. For audio-video generation, we conduct anonymized pairwise human comparisons of audio-visual quality and prompt alignment (Appendix D.2).

4.1 Main Results

Quantitative Results. Tab. 1 compares VA-TDM with few-step baselines and their teachers on text-to-image generation. At four NFE, VA-TDM improves compositional prompt following and text rendering on both backbones. On SD3.5-Medium, it leads the compared few-step models on all verifiable, quality, and preference metrics, raising GenEval from the strongest baseline TDM's 0.61 to 0.67 while increasing TDM's diversity. On FLUX2, gains in prompt following and most quality and preference metrics are accompanied by an LPIPS increase from 0.47 to 0.56. Together, the results demonstrate the state-of-the-art performance of our proposed VA-TDM in few-step generation.

Qualitative Results. Across visual comparisons in images and videos (Figs. 1 and 4), VA-TDM consistently delivers better visual quality compared with few-step baselines, notably approaching the many-step teacher. ***Moreover, we find that previous methods either suffer from oversaturation or excessively dark outputs, whereas VA-TDM effectively avoids these undesirable artifacts.***

4.2 Ablation Study

We examine the components introduced in Sec. 3.3 using four-step SD3.5-Medium (Tab. 2). Following Eq. (22), λ_m mixes the A-form and B-form games; the separate coefficient λ in Eq. (20) controls the path/flow-product mixture within each form, and λ_r weights the online field's regularization toward the teacher velocity. Starting from $\lambda_m = 0.8$, the final three rows report cumulative configurations with generator-update mixing, field regularization, and critic correction.

Dual-Form Velocity Mixing. The A-form rectified velocity and B-form Gaussian velocity in Sec. 3.1 yield complementary results. A-form alone improves GenEval over B-form alone, but has lower OCR and HPS scores. Combining the two games produces higher OCR and ImageReward than either endpoint while retaining the A-form diversity. This supports combining their supervision.

Shared Backbone. Sharing the online-field backbone between the two forms not only reduces training cost but also improves performance. We believe that this is because the high-variance B-

Table 2: **Ablation study.** Best / second-best results are **bold** / underlined.

Configuration	Verifiable		ImgQ	Preference				Div.
	GenEval	OCR	Aest.	ImgRwd	PickScore	HPSv2	HPSv3	LPIPS
λ_m=1 (A-form only)	0.64	0.55	5.53	1.00	0.8671	0.2949	11.20	0.58
λ_m=0 (B-form only)	0.61	0.57	5.57	0.99	0.8665	0.2987	11.47	**0.59**
λ_m=0 w/o transition-aware field	0.58	0.53	5.56	0.99	0.8655	0.2964	11.13	0.56
λ_m=0.8 w/o shared backbone	0.62	0.56	5.55	0.99	0.8669	0.2971	11.33	0.57
λ_m=0.8	0.63	0.58	5.56	1.01	0.8674	0.2983	11.47	0.58
+ path/flow-product mixing (λ=0.7)	0.65	0.57	5.56	0.96	0.8687	0.2950	11.39	**0.59**
+ field regularization (λ_r=1)	0.66	0.58	5.58	0.96	0.8682	0.2952	11.44	0.57
+ regularization-corrected critic	**0.67**	**0.59**	**5.65**	**1.02**	**0.8702**	**0.3010**	**11.89**	0.58

form is difficult to learn on its own, whereas the shared backbone allows the network to learn from both online fields jointly, which potentially leads to better internal representations and thus provides better supervision signals.

Path/Flow-Product Mixing. The mixed update notably improves GenEval and diversity, while the other metrics remain roughly unchanged. This indicates that the mixed update allows us to find a better-performing sweet spot.

Online Field Regularization and Regularization-Corrected Critic. We augment the field regression in Eq. (21) with the teacher-velocity regularizer, using $\lambda_r = 1$. Even without critic correction, this also slightly improves GenEval, OCR, supporting the benefit of retaining teacher information during online field learning. By applying the regularization-corrected critic $\mathbf{f}_\psi^F = (1+\lambda_r)(\mathbf{u}_\tau - \mathbf{h}_\psi^F)$ and adjusted loss coefficients, the resulting configuration notably improves the performance and achieves the strongest verifiable, quality, and preference scores among variants.

Transition-Aware Field Estimation. This enables accurate tracking of the student field and notably improves instruction following (GenEval), text rendering (OCR), and diversity (LPIPS).

5 Related Works

Diffusion Distillation. Distilled models generate images from noise in one or a few steps. Two main approaches transfer knowledge from pre-trained DMs: 1) Trajectory distillation (Luhman & Luhman, 2021; Meng et al., 2023; Salimans & Ho, 2022; Song et al., 2023; Song & Dhariwal, 2024; Yan et al., 2024) matches teacher ODE trajectories at the instance level, facing challenging instance-wise alignment and numerical errors in teacher PF-ODE solutions; 2) Distribution matching via score distillation (Yin et al., 2023; Luo et al., 2023; Zhou et al., 2024b;a; Wu et al., 2026; Liu et al., 2026) aligns student and teacher distributions using objectives such as reverse KL, Fisher divergence, and moment matching. FGM aligns velocity fields but considers only Gaussian forward processes Huang et al. (2024), where its objective reduces to Fisher divergence (Zhou et al., 2024b). TDM (Luo et al., 2025) has shown strong few-step performance by learning partial denoising at each step. Our VA-TDM unifies score distillation, Fisher divergence, and moment matching while yielding new distillation objectives. Combined with transition-aware online field learning, it achieves state-of-the-art few-step performance. Additional related work is discussed in Appendix C.

6 Discussion

We introduced Velocity Adversarial Distillation (VAD), a unified framework that guides distillation through velocity fields constructed from the student's sampling process. This perspective is particularly timely as flow matching becomes increasingly prevalent in generative pretraining, while distillation still often relies on score-based objectives via the score–velocity correspondence under Gaussian probability paths. By formulating distillation directly in terms of velocity fields, VAD aligns distillation more closely with the flow matching paradigm and offers greater flexibility in designing supervision. Building on this framework, VA-TDM combines dual-form supervision with transition-aware online velocity learning to supervise partial denoising along few-step student trajectories, without adding inference cost. Experiments across text-to-image/video generation demonstrate that VA-TDM consistently achieves state-of-the-art four-step performance. VA-TDM also scales effectively to 33B MiniMax-H3, outperforming Fast-H3 in human evaluation and substantially narrowing the perceptual quality gap to the many-step teacher. These results highlight the value of our work, establishing VAD as a principled foundation for designing effective few-step distillation methods.

AI USE STATEMENT

We used generative AI tools to assist with interpreting results, and drafting and revising the manuscript. AI assistance also supported literature searches and the development and debugging of research code. The authors retain responsibility for the research decisions and for verifying the correctness, originality, and reproducibility of the work, including all AI-assisted text, theoretical claims, code, and empirical analyses.

ETHICS STATEMENT

This work studies efficient image and video generation using existing benchmarks and pretrained diffusion teachers. Distilled models can inherit biases, memorized training samples, and potentially harmful content from their teachers, which are all popular open-source models.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our findings, our complete code and experimental setup will be made publicly available before publication.

REFERENCES

Black Forest Labs. FLUX.2 [klein]. https://bfl.ai/models/flux-2-klein, 2026.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

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

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Zemin Huang, Zhengyang Geng, Weijian Luo, and Guo jun Qi. Flow generator matching, 2024. URL https://arxiv.org/abs/2410.19310.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34: 14745–14758, 2021.

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, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

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.

Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021.

Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi. Self-guidance: Boosting flow and diffusion generation on their own. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Adversarial flow models, 2026. URL https://arxiv.org/abs/2511.22475.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, and Hongsheng Li. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jBztvOiCKE.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

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. *Advances in Neural Information Processing Systems*, 36, 2023.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ogk236hsJM.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. *arXiv preprint arXiv:2503.06674*, 2025.

Jiajun Ma, Shuchen Xue, Tianyang Hu, Wenjia Wang, Zhaoqiang Liu, Zhenguo Li, Zhi-Ming Ma, and Kenji Kawaguchi. The surprising effectiveness of skip-tuning in diffusion sampling. *arXiv preprint arXiv:2402.15170*, 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 (CVPR)*, pp. 14297–14306, June 2023.

MiniMax. MiniMax-H3. Hugging Face model card, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3. Accessed: 2026-09-26.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching, 2024. URL https://arxiv.org/abs/2406.04103.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4733–4743, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Shuchen Xue, Zhaoqiang Liu, Fei Chen, Shifeng Zhang, Tianyang Hu, Enze Xie, and Zhenguo Li. Accelerating diffusion sampling with optimized time steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8292–8301, 2024.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models, 2025. URL https://arxiv.org/abs/2509.25050.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis, 2024. URL https://arxiv.org/abs/2405.14867.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step, 2024a. URL `https://arxiv.org/abs/2410.14919`.

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.

Jiapeng Zhu, Ceyuan Yang, Kecheng Zheng, Yinghao Xu, Zifan Shi, Yifei Zhang, Qifeng Chen, and Yujun Shen. Exploring sparse moe in gans for text-conditioned image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18411–18423, 2025.

A Proofs of Main Results

Throughout, time variables and conditioning information are fixed. For the population game in Eq. (12), assume that the teacher field and sampled target have finite second moments, and that critic optimization ranges over all square-integrable fields under the student's marginal law. Proposition 3 additionally assumes realizability: there exists a student $\theta^\star$ with $\bar{\mathbf{v}}^F_{\theta^\star,\tau} = \mathbf{u}_\tau$ almost everywhere under $p_{\theta^\star,\tau}$. The critic maximizes V_F, while the student minimizes $\mathcal{L}^F_{\mathrm{path}}$ with the critic fixed.

A.1 Proof of Lemma 2

Proof. Conditioning on $\mathbf{x}_\tau$ gives the marginal payoff

$$V_F(\mathbf{f},\theta) = \mathbb{E}_{p_{\theta,\tau}}\left[\mathbf{f}^\top\left(\mathbf{u}_\tau - \bar{\mathbf{v}}^F_{\theta,\tau}\right) - \frac{1}{2}\|\mathbf{f}\|^2\right]. \tag{23}$$

Writing $\mathbf{r}_\theta = \mathbf{u}_\tau - \bar{\mathbf{v}}^F_{\theta,\tau}$ and completing the square yields

$$V_F(\mathbf{f},\theta) = \frac{1}{2}\mathbb{E}_{p_{\theta,\tau}}\|\mathbf{r}_\theta\|^2 - \frac{1}{2}\mathbb{E}_{p_{\theta,\tau}}\|\mathbf{f}-\mathbf{r}_\theta\|^2. \tag{24}$$

The upper bound is attained exactly when $\mathbf{f} = \mathbf{r}_\theta$ almost everywhere under $p_{\theta,\tau}$, proving the optimal critic and value. □

Parameterizing $\mathbf{f}^F_\psi = \mathbf{u}_\tau - \mathbf{h}^F_\psi$ makes critic optimization equivalent to the field regression in Eq. (21), by the conditional variance decomposition.

A.2 Proof of Proposition 3

Proof. Choose a realizable matching student $\theta^\star$ and the globally zero critic $\mathbf{f}^\star \equiv \mathbf{0}$. For any admissible critic,

$$V_F(\mathbf{f},\theta^\star) = -\frac{1}{2}\mathbb{E}_{p_{\theta^\star,\tau}}\|\mathbf{f}\|^2 \le 0 = V_F(\mathbf{f}^\star,\theta^\star), \tag{25}$$

so $\mathbf{f}^\star$ is a best response. With this critic, $\mathcal{L}^F_{\mathrm{path}}(\theta;\mathbf{f}^\star) = 0$ for every student, so $\theta^\star$ is also a best response. Thus $(\theta^\star,\mathbf{f}^\star)$ is a Nash equilibrium. By Lemma 2,

$$J_F(\theta) \coloneqq \max_{\mathbf{f}} V_F(\mathbf{f},\theta) = \frac{1}{2}\mathbb{E}_{p_{\theta,\tau}}\|\mathbf{u}_\tau - \bar{\mathbf{v}}^F_{\theta,\tau}\|^2 \ge 0, \tag{26}$$

and $J_F(\theta^\star) = 0$, establishing the claimed optimality. □

A.3 Proof of Lemma 4

Our proof uses the conditional-projection principle underlying denoising score matching (Vincent, 2011) and the score-projection identity used in SiD (Zhou et al., 2024b) and SIM (Luo et al., 2024). FGM (Huang et al., 2024) develops the corresponding flow-product identity. Here the argument applies to conditional target fields under the student's composite sampling law, without requiring a Gaussian forward kernel or a continuity-equation interpretation of the target field.

Assume differentiable marginal densities and conditional fields near θ_0, integrability, and domination sufficient to interchange differentiation and integration. Reparameterized derivatives are assumed to exist and commute with expectation. Input laws and time variables are held fixed during differentiation; auxiliary parameters are frozen while their input derivatives are retained.

Proof. Let P_θ be the joint sampling law, write $X = \mathbf{x}_\tau$ and $V_\theta^F = \mathbf{v}_\tau^F(\mathbf{x}_\tau \mid \mathbf{x}_s, \hat{\mathbf{x}}_0)$, and define $\bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}) = \mathbb{E}_{P_\theta}[V_\theta^F \mid X = \mathbf{x}]$. For any admissible test field $\mathbf{a}_\theta$, conditional expectation gives

$$\mathbb{E}_{P_\theta}\left[\mathbf{a}_\theta(X)^\top \left(V_\theta^F - \bar{\mathbf{v}}_{\theta,\tau}^F(X)\right)\right] = 0. \tag{27}$$

Fix θ_0, set $\mathbf{h}_0^F = \bar{\mathbf{v}}_{\theta_0,\tau}^F$, and hold $\mathbf{f}_0$ and $\mathbf{h}_0^F$ fixed as functions. The projection identity gives

$$H(\theta;\theta_0) = \mathbb{E}_{P_\theta}\left[\mathbf{f}_0(X)^\top \left(V_\theta^F - \mathbf{h}_0^F(X)\right)\right] \tag{28}$$

$$= \int p_{\theta,\tau}(\mathbf{x})\, \mathbf{f}_0(\mathbf{x})^\top \left(\bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x}) - \mathbf{h}_0^F(\mathbf{x})\right) \mathrm{d}\mathbf{x}. \tag{29}$$

At $\theta = \theta_0$, the centered difference vanishes pointwise, so its product with the sampling-density derivative is zero. Hence

$$\nabla_\theta H(\theta;\theta_0)|_{\theta_0} = \mathbb{E}_{p_{\theta_0,\tau}}\left[\mathbf{f}_0(\mathbf{x})^\top\, \partial_\theta \bar{\mathbf{v}}_{\theta,\tau}^F(\mathbf{x})\big|_{\theta_0}\right], \tag{30}$$

where ∂_θ differentiates at a fixed spatial argument. Since $\mathcal{G}_{\mathrm{fp}}^F = -H$, this is precisely

$$\nabla_\theta \mathcal{G}_{\mathrm{fp}}^F\big|_{\theta_0} = \nabla_\theta \mathcal{G}_{\mathrm{mean}}^F(\theta;\theta_0,\mathbf{f}_0)\big|_{\theta_0}, \tag{31}$$

whose right-hand side holds the outer measure fixed at $p_{\theta_0,\tau}$. The argument allows both the target and the joint law to depend on θ, and applies to both A-form and B-form. □

The identity uses the exact conditional mean; a learned auxiliary field provides an approximation. The practical mixed update in Eq. (20) additionally detaches the critic output for stability while retaining input gradients through the centering field; its gradient is not identified with the exact local gradient above.

A.4 Proof of Theorem 5

Fix $0 \le s < t \le 1$, $s < \tau < 1$, and $\eta = 1$. Let P_θ denote the law of the local reparameterized chain

$$\hat{\mathbf{x}}_0 = g_\theta(\mathbf{x}_t, t), \qquad \mathbf{x}_s = \frac{s}{t}\mathbf{x}_t + \left(1 - \frac{s}{t}\right)\hat{\mathbf{x}}_0, \qquad \mathbf{x}_\tau = \alpha_{\tau|s}\mathbf{x}_s + \beta_{\tau|s}\boldsymbol{\varepsilon}, \tag{32}$$

with fixed input law and independent standard Gaussian $\boldsymbol{\varepsilon}$. Use the regularity conditions of Appendix A.3, an exact teacher, and finite reverse KL with the boundary conditions needed for its pathwise gradient. Write D_θ for total sample derivatives with base noises fixed, and ∂_θ for field derivatives at fixed spatial arguments. All gradient identities are evaluated at a reference student $\bar{\theta}$, using the exact reference field $\bar{\mathbf{v}}_{\bar{\theta},\tau}^F$ and critic $\mathbf{u}_\tau - \bar{\mathbf{v}}_{\bar{\theta},\tau}^F$. For a teacher-regularized regressor, the unregularized reference field is $(1+\lambda_r)\mathbf{h}_{\psi^\star}^F - \lambda_r \mathbf{u}_\tau$.

Proof. **(i) Matching.** For A-form, conditioning the target on $\mathbf{x}_\tau$ gives

$$\bar{\mathbf{v}}_{\theta,\tau}^A(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{m}_\theta(\mathbf{x})}{\tau}, \qquad \mathbf{f}_A^\star(\mathbf{x}) = \frac{\mathbf{m}_\theta(\mathbf{x}) - \mathbf{m}_q(\mathbf{x})}{\tau}, \tag{33}$$

with the moments defined in Eq. (18). Thus zero residual is equivalent to conditional-moment matching. The sampled target also has the endpoint-path interpretation: fixing $\boldsymbol{\varepsilon}_{\mathrm{eff}} = [\mathbf{x}_\tau - (1-\tau)\hat{\mathbf{x}}_0]/\tau$, the interpolation $\mathbf{y}_r = (1-r)\hat{\mathbf{x}}_0 + r\boldsymbol{\varepsilon}_{\mathrm{eff}}$ satisfies $\mathbf{y}_\tau = \mathbf{x}_\tau$ and $\partial_r \mathbf{y}_r = \mathbf{v}_\tau^A$.

For B-form, Gaussian differentiation gives $\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{x}_\tau = \mathbf{x}] = -\beta_{\tau|s}\nabla \log p_{\theta,\tau}(\mathbf{x})$. Using the coefficients of Eq. (5), with $\alpha = \alpha_{\tau|s}$ and $\beta = \beta_{\tau|s}$, yields

$$\begin{aligned}\bar{\mathbf{v}}^B_{\theta,\tau}(\mathbf{x}) &= \frac{\dot\alpha}{\alpha}\mathbf{x} - \left(\beta\dot\beta - \frac{\dot\alpha}{\alpha}\beta^2\right)\nabla \log p_{\theta,\tau}(\mathbf{x}) \\ &= -\frac{\mathbf{x}}{1-\tau} - \frac{\tau}{1-\tau}\nabla \log p_{\theta,\tau}(\mathbf{x}), \\ \mathbf{f}^\star_B &= \frac{\tau}{1-\tau}(\nabla \log p_{\theta,\tau} - \nabla \log q_\tau).\end{aligned} \tag{34}$$

The kernel maps q_s to q_τ, since $\alpha_{\tau|s}(1-s) = 1-\tau$ and $\alpha^2_{\tau|s}s^2 + \beta^2_{\tau|s} = \tau^2$. Gaussian smoothing gives positive densities on $\mathbb{R}^d$, so equal scores imply equal normalized densities. Since $\alpha_{\tau|s} > 0$ and the Gaussian characteristic function never vanishes, deconvolution makes $p_{\theta,\tau} = q_\tau$ equivalent to $p_{\theta,s} = q_s$.

(ii) The clean endpoint. At $s = 0$, $\mathbf{x}_s = \hat{\mathbf{x}}_0$ and $\mathbf{x}_\tau = (1-\tau)\hat{\mathbf{x}}_0 + \tau\mathbf{z}_\tau$, giving

$$\mathbf{v}^A_\tau = \frac{\mathbf{x}_\tau - \hat{\mathbf{x}}_0}{\tau} = \mathbf{z}_\tau - \hat{\mathbf{x}}_0 = \mathbf{v}^B_\tau. \tag{35}$$

The sampled targets coincide, so their conditional means, optimal critics, and gradients under identical differentiation conventions coincide.

(iii) Pathwise updates with the critic output detached. Set $a = 1 - s/t > 0$ and $c = \alpha_{\tau|s}a \in (0,1)$. From Eq. (32),

$$\mathrm{D}_\theta\mathbf{x}_s = a\,\mathrm{D}_\theta\hat{\mathbf{x}}_0, \qquad \mathrm{D}_\theta\mathbf{x}_\tau = c\,\mathrm{D}_\theta\hat{\mathbf{x}}_0, \tag{36}$$

and the sampled targets satisfy

$$\mathrm{D}_\theta\mathbf{v}^A_\tau = -\frac{1-c}{\tau}\mathrm{D}_\theta\hat{\mathbf{x}}_0, \qquad \mathrm{D}_\theta\mathbf{v}^B_\tau = -\frac{a}{1-s}\mathrm{D}_\theta\hat{\mathbf{x}}_0. \tag{37}$$

The reverse-KL pathwise identity is

$$\nabla_\theta \mathrm{KL}(p_{\theta,\tau}\|q_\tau) = \mathbb{E}\left[(\mathrm{D}_\theta\mathbf{x}_\tau)^\top(\nabla \log p_{\theta,\tau} - \nabla \log q_\tau)(\mathbf{x}_\tau)\right]. \tag{38}$$

Substituting the optimal critics into Eq. (19) gives

$$\nabla_\theta\mathcal{L}^A_{\mathrm{path,sg}} = \frac{1-c}{\tau^2}\mathbb{E}\left[(\mathrm{D}_\theta\hat{\mathbf{x}}_0)^\top(\mathbf{m}_\theta - \mathbf{m}_q)(\mathbf{x}_\tau)\right], \tag{39}$$

$$\nabla_\theta\mathcal{L}^B_{\mathrm{path,sg}} = \frac{a}{1-s}\mathbb{E}[(\mathrm{D}_\theta\hat{\mathbf{x}}_0)^\top\mathbf{f}^\star_B] = \frac{\tau}{(1-\tau)^2}\nabla_\theta\mathrm{KL}(p_{\theta,\tau}\|q_\tau). \tag{40}$$

The A-form expectation is the MMD surrogate update made explicit below; the B-form gives reverse KL. Both coefficients are positive time weights.

(iv) Flow-product updates without output detachment. Define the discrepancy with its outer measure held fixed:

$$\mathcal{D}_F(\theta;\bar\theta) = \frac{1}{2}\mathbb{E}_{p_{\bar\theta,\tau}}\|\mathbf{u}_\tau - \bar{\mathbf{v}}^F_{\theta,\tau}\|^2. \tag{41}$$

Lemma 4, applied to the optimal reference critic, gives

$$\nabla_\theta\mathcal{G}^F_{\mathrm{fp}}\big|_{\bar\theta} = -\mathbb{E}_{p_{\bar\theta,\tau}}\left[\left(\partial_\theta\bar{\mathbf{v}}^F_{\theta,\tau}\big|_{\bar\theta}\right)^\top\mathbf{f}^\star_F\right] = \nabla_\theta\mathcal{D}_F(\theta;\bar\theta)\big|_{\bar\theta}. \tag{42}$$

The residual identities in part (i) identify the objectives as

$$\mathcal{D}_A(\theta;\bar\theta) = \frac{1}{2\tau^2}\mathbb{E}_{p_{\bar\theta,\tau}}\|\mathbf{m}_\theta - \mathbf{m}_q\|^2, \tag{43}$$

$$\mathcal{D}_B(\theta;\bar\theta) = \frac{\tau^2}{2(1-\tau)^2}\mathbb{E}_{p_{\bar\theta,\tau}}\|\nabla \log p_{\theta,\tau} - \nabla \log q_\tau\|^2. \tag{44}$$

These are the conditional-moment and scaled Fisher discrepancies, respectively, with the fixed-outer-measure gradient convention.

MMD objective and surrogate. To make the MMD correspondence explicit (Salimans et al., 2024), write $\boldsymbol{\delta}_\theta = \mathbf{m}_\theta - \mathbf{m}_q$. Its moment objective and local fixed-measure version are

$$J_{\mathrm{MM}}(\theta) = \frac{1}{2}\mathbb{E}_{p_{\theta,\tau}}\|\boldsymbol{\delta}_\theta\|^2, \qquad L_{\mathrm{MM}}(\theta;\bar{\theta}) = \frac{1}{2}\mathbb{E}_{p_{\bar{\theta},\tau}}\|\boldsymbol{\delta}_\theta\|^2 = \tau^2\mathcal{D}_A(\theta;\bar{\theta}). \tag{45}$$

The exact local gradient is

$$\mathbf{g}_{\mathrm{MM}} = \nabla_\theta L_{\mathrm{MM}}|_{\bar{\theta}} = \mathbb{E}_{p_{\bar{\theta},\tau}}\left[\left(\partial_\theta \mathbf{m}_\theta|_{\bar{\theta}}\right)^\top \boldsymbol{\delta}_{\bar{\theta}}\right]. \tag{46}$$

MMD approximates the implicit conditional-mean derivative by the sampled clean-prediction derivative, giving

$$\mathbf{g}_{\mathrm{MM}} \approx \mathbf{g}_{\mathrm{sur}} := \mathbb{E}_{P_{\bar{\theta}}}\left[\left(\mathrm{D}_\theta \hat{\mathbf{x}}_0|_{\bar{\theta}}\right)^\top \boldsymbol{\delta}_{\bar{\theta}}(\mathbf{x}_\tau)\right]. \tag{47}$$

This is the gradient of the detached surrogate

$$L_{\mathrm{sur}}(\theta) = \mathbb{E}_{P_\theta}\left[\hat{\mathbf{x}}_0^\top \,\mathrm{sg}(\boldsymbol{\delta}_\theta(\mathbf{x}_\tau))\right]. \tag{48}$$

Parts (iii)–(iv) therefore give both correspondences directly:

$$\tau^2\,\nabla_\theta \mathcal{G}^A_{\mathrm{fp}}\big|_{\bar{\theta}} = \mathbf{g}_{\mathrm{MM}}, \tag{49}$$

$$\nabla_\theta \mathcal{L}^A_{\mathrm{path,sg}}\big|_{\bar{\theta}} = \frac{1-c}{\tau^2}\mathbf{g}_{\mathrm{sur}}. \tag{50}$$

Thus flow-product recovers the exact local moment gradient, while the detached pathwise update recovers MMD's surrogate direction up to time weighting. □

B Supporting Derivations

B.1 Reference DSM Variance

Consider the reference Gaussian forward process with $0 < s < \tau < 1$, $\mathbf{x}_s = (1-s)\mathbf{x}_0 + s\mathbf{e}_s$, and $\mathbf{x}_\tau = \alpha_{\tau|s}\mathbf{x}_s + \beta_{\tau|s}\mathbf{e}_\tau$, where the standard Gaussian noises are independent of each other and $\mathbf{x}_0$. The two DSM targets

$$\mathbf{S}_{s\to\tau} = -\frac{\mathbf{x}_\tau - \alpha_{\tau|s}\mathbf{x}_s}{\beta^2_{\tau|s}}, \qquad \mathbf{S}_{0\to\tau} = -\frac{\mathbf{x}_\tau - (1-\tau)\mathbf{x}_0}{\tau^2} \tag{51}$$

have the same conditional mean $\nabla \log q_\tau(\mathbf{x}_\tau)$. The Gaussian bridge and the law of total covariance give the comparison of Xue et al. (2025) in our notation:

$$\mathrm{Cov}(\mathbf{S}_{s\to\tau} \mid \mathbf{x}_\tau) = \mathrm{Cov}(\mathbf{S}_{0\to\tau} \mid \mathbf{x}_\tau) + \left(\frac{1}{\beta^2_{\tau|s}} - \frac{1}{\tau^2}\right)\mathbf{I}. \tag{52}$$

The excess is positive for $0 < s < \tau < 1$ and vanishes at $s = 0$. This reference-process comparison motivates the noisy-target DSM discussion in Sec. 3.1.

B.2 Clean-Target and Path-Marginal Score Mismatch

Fix t, s, τ with $s < \tau < 1$ and the input law. Let $R_\theta(\mathrm{d}\mathbf{x}_s, \mathrm{d}\hat{\mathbf{x}}_0)$ be the joint law of the intermediate output and clean anchor. The two perturbed marginals are

$$\begin{aligned} q(\mathbf{x}_\tau \mid \hat{\mathbf{x}}_0) &= \mathcal{N}(\mathbf{x}_\tau; (1-\tau)\hat{\mathbf{x}}_0, \tau^2\mathbf{I}), \\ p^{\mathrm{clean}}_{\theta,\tau}(\mathbf{x}_\tau) &= \int q(\mathbf{x}_\tau \mid \hat{\mathbf{x}}_0) R_\theta(\mathrm{d}\mathbf{x}_s, \mathrm{d}\hat{\mathbf{x}}_0), \\ p^{\mathrm{path}}_{\theta,\tau}(\mathbf{x}_\tau) &= \int q(\mathbf{x}_\tau \mid \mathbf{x}_s) R_\theta(\mathrm{d}\mathbf{x}_s, \mathrm{d}\hat{\mathbf{x}}_0). \end{aligned} \tag{53}$$

TDM's ideal importance-weighted clean-target objective (Luo et al., 2025) uses the ratio of these conditional kernels. Since both kernels have positive densities, for any integrable loss ℓ_ψ,

$$w(\mathbf{x}_\tau, \mathbf{x}_s, \hat{\mathbf{x}}_0) = \frac{q(\mathbf{x}_\tau \mid \hat{\mathbf{x}}_0)}{q(\mathbf{x}_\tau \mid \mathbf{x}_s)}, \tag{54}$$

$$\mathbb{E}_{R_\theta q(\mathbf{x}_\tau|\mathbf{x}_s)}[w\ell_\psi] = \mathbb{E}_{R_\theta q(\mathbf{x}_\tau|\hat{\mathbf{x}}_0)}[\ell_\psi]. \tag{55}$$

For $\ell_\psi = \omega_\tau \|h_\psi(\mathbf{x}_\tau, \tau) - \hat{\mathbf{x}}_0\|^2$ with $\omega_\tau > 0$, an unrestricted regressor with finite second moments has optimum $h^*(\mathbf{x}_\tau, \tau) = \mathbb{E}_{R_\theta q(\mathbf{x}_\tau|\hat{\mathbf{x}}_0)}[\hat{\mathbf{x}}_0 \mid \mathbf{x}_\tau]$. Gaussian differentiation identifies its score and the required path score as

$$\begin{aligned} \nabla \log p^{\text{clean}}_{\theta,\tau}(\mathbf{x}_\tau) &= -\frac{\mathbf{x}_\tau - (1-\tau)h^*(\mathbf{x}_\tau, \tau)}{\tau^2}, \\ \nabla \log p^{\text{path}}_{\theta,\tau}(\mathbf{x}_\tau) &= \mathbb{E}_{R_\theta q(\mathbf{x}_\tau|\mathbf{x}_s)}\left[-\frac{\mathbf{x}_\tau - \alpha_{\tau|s}\mathbf{x}_s}{\beta^2_{\tau|s}} \,\middle|\, \mathbf{x}_\tau \right]. \end{aligned} \tag{56}$$

Thus importance weighting recovers the clean-denoising objective, whose population score can differ from the path score needed by the generator. For example, in one dimension with $\hat{x}_0 = \hat{\epsilon} = 0$ and $\eta = 1$, the transition gives $x_s = 0$: the path marginal is $\mathcal{N}(0, \beta^2_{\tau|s})$ and the clean marginal is $\mathcal{N}(0, \tau^2)$, with different variances for $s > 0$. At $s = 0$, $\mathbf{x}_s = \hat{\mathbf{x}}_0$ and the two marginals coincide.

C Additional Related Works

Training-Free Diffusion Acceleration. Training-free methods for accelerating diffusion model sampling have progressed substantially (Lu et al., 2022; Zhao et al., 2023; Xue et al., 2024; Si et al., 2024; Ma et al., 2024; Li et al., 2025). However, diffusion distillation remains essential for high-quality few-step generation.

Generative adversarial networks.. Early GAN research achieved strong results on domain-specific datasets (Reed et al., 2016; Zhang et al., 2017; Karras et al., 2020; 2021), while it has been effectively extended these successes to ImageNet (Brock et al., 2019; Sauer et al., 2022; Russakovsky et al., 2015). However, training instability and limited scalability remain key challenges. Large-scale text-to-image GANs still rely on convolutional architectures with complex designs (Kang et al., 2023; Zhu et al., 2025), and scaling transformer-based GANs remains difficult (Jiang et al., 2021; Lee et al., 2021). Adversarial Flow Models (AFM) (Lin et al., 2026) train many-step flow-matching models through an adversarial game in velocity space, requiring discriminator Jacobian-vector products (JVPs). Our work also formulates an adversarial game in velocity space, but targets few-step distillation and avoids expensive JVP computation.

D Experimental Details

This section provides the training details for the experiments in Sec. 4.

D.1 Training Configurations

VA-TDM trains a generator and an online auxiliary field while keeping the teacher fixed. The generator combines the two velocity forms through Eq. (22), and the auxiliary field follows the regression and regularization objectives in Sec. 3.3. By default, we apply LoRA fine-tuning with rank 64. We update the online score and generator once each iteration. We use the AdamW optimizer and set $(\beta_1, \beta_2) = (0., 0.95)$.

D.2 User Study

The study in Fig. 2 compares MiniMax-H3 with Fast-H3, VA-TDM with MiniMax-H3, and VA-TDM with Fast-H3. We carried out a user study in which participants were shown two anonymized audio-videos, each produced by a different model, and were asked to pick the one with superior audio-visual quality and closer alignment with the prompt. For audio-video generation, we randomly chose 10 prompts, and every generated audio-video was manually checked to confirm it contained no inappropriate or harmful content.

E Additional Visual Samples

Figs. 5 to 9 provide additional text-to-video comparisons complementing Fig. 1. Each figure presents six selected frames per method, with the summarized prompt shown above the panels. From top to

bottom, the rows correspond to MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE). The examples cover human–animal interactions, clothing changes, fantasy combat, stylized game scenes, and multi-scene cinematic sequences.



Figure 5: Additional visual comparison of MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE).

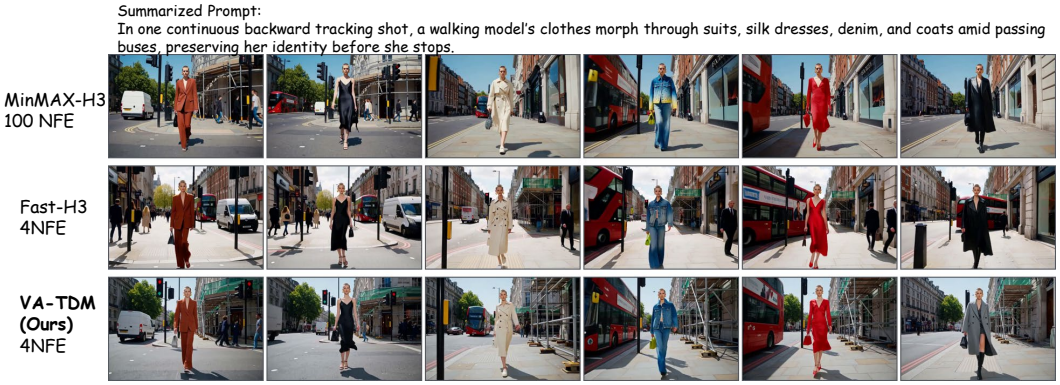

Figure 6: Additional visual comparison of MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE).



Figure 7: Additional visual comparison of MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE).



Figure 8: Additional visual comparison of MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE).

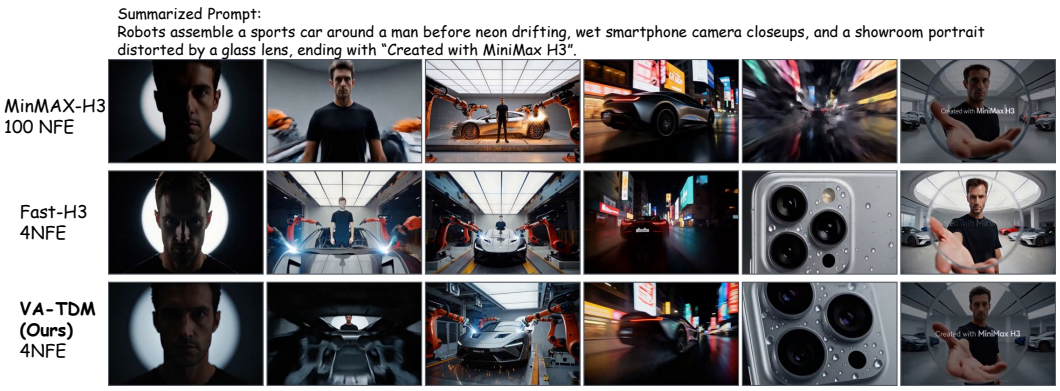

Figure 9: Additional visual comparison of MinMAX-H3 (100 NFE), Fast-H3 (4 NFE), and VA-TDM-H3 (4 NFE).