sDMD: Sinkhorn Distribution Matching Distillation

Anonymous authors
Paper under double-blind review

Abstract

Distribution Matching Distillation (DMD) distills multi-step diffusion models into few-step generators by minimizing reverse KL divergence. Despite its strong performance, DMD guides samples by local density alone, which introduces two limitations: teacher scores can be unreliable in regions of low teacher density, and the mode-seeking behavior of reverse KL can collapse the student onto a subset of the teacher modes. We first connect DMD to Wasserstein gradient flow, showing that its gradient follows from a fixed-point formulation with integral KL (IKL) as the energy functional. This perspective motivates augmenting IKL with an optimal transport term that addresses these limitations by guiding outlier samples toward the target and promoting mode coverage. We realize this term using Sinkhorn divergence estimated from finite batches. In practice, however, errors in the estimated Sinkhorn velocity can perturb in-domain samples for which IKL guidance is already accurate. To this end, we introduce `sDMD`, which adaptively activates the Sinkhorn correction based on discriminator outputs, focusing on outlier samples where IKL may be unreliable. With 2–4 sampling steps on FLUX.1-dev, Qwen-Image-20B, and Wan2.1-14B, `sDMD` outperforms distillation baselines on GenEval, DPG, WISE, and VBench while improving sample diversity over DMD2.

1 Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2021; Ho et al., 2020) generate high-quality images and videos but require many sampling steps. Diffusion distillation reduces this cost by transferring knowledge from a pretrained teacher to a student that generates samples in one or a few steps. Existing approaches have advanced few-step generation by learning mappings along teacher trajectories (Geng et al., 2025; Song et al., 2023), using adversarial supervision (Sauer et al., 2024; Lin et al., 2025), or directly matching student and teacher distributions (Xu et al., 2025; Zhou et al., 2024). Among these, a representative approach is DMD (Yin et al., 2024b), which optimizes a reverse KL objective using the difference between teacher and student scores. It has been widely adopted due to its high sample fidelity and scalability in large-scale image and video generation (Fan et al., 2025; Zheng et al., 2026b; Ge et al., 2025).

To better understand and improve the training dynamics of DMD, we view diffusion distillation from the perspective of Wasserstein gradient flow (WGF) (Ambrosio et al., 2008). WGF evolves a probability distribution by steepest descent of an energy functional in Wasserstein space, with the corresponding velocity field specifying the update direction for each sample. While recent works (Han et al., 2026; He et al., 2026) use WGF to train one-step generators from scratch, we establish its connection to diffusion distillation by recovering the DMD gradient through fixed-point regression under the integral KL (IKL) energy, defined as reverse KL averaged over diffusion noise levels (Sec. 4.1). This identifies the energy underlying DMD and the velocity field that guides generated samples, allowing us to analyze the behavior and limitations of DMD and explore complementary energy functionals to improve distillation.

The IKL velocity of DMD reveals two practical limitations (Fig. 1). First, learned teacher scores can be unreliable for outlier samples, and the resulting updates may fail to guide them toward target modes. Although adding diffusion noise can improve distributional overlap, outliers may remain poorly supported at low noise levels. Higher noise levels can improve their support but blur the local structure of target modes. Second, the mode-seeking behavior of reverse KL can concentrate the student on a subset of target modes, potentially reducing the diversity of generated samples.

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.

To address these limitations, we augment IKL with an optimal transport (OT) term. OT couples student and target samples globally, providing robust guidance for outlier samples and promoting mode coverage without relying on learned teacher scores. To exploit their complementary strengths, we propose combining global OT guidance with local IKL score updates that help capture fine details. Our analysis shows that sufficiently strong exact OT regularization ensures convergence of the combined flow to the target under regularity assumptions (Sec. 4.2). For efficient computation, we use Sinkhorn divergence, a debiased form of entropic OT. While Sinkhorn does not automatically inherit this guarantee, we establish sufficient conditions on its approximation error and regularization strength for the combined flow to converge to the target (Sec. 5).

In practice, the transport maps used to compute Sinkhorn updates are estimated from finite samples, with larger batches generally reducing estimation error at greater computational cost (Genevay et al., 2019). Notably, our experiments show that even with relatively few samples, Sinkhorn can provide more informative guidance than IKL for outlier samples where learned teacher scores are unreliable. However, for well-supported samples, estimation errors in Sinkhorn can instead interfere with the accurate guidance already provided by IKL. To this end, we introduce `sDMD`, which adaptively activates Sinkhorn guidance for samples with relatively weak teacher support based on discriminator outputs. We reuse the discriminator trained for adversarial regularization in DMD2 (Yin et al., 2024a), without training a separate model.

We evaluate `sDMD` using 2–4 sampling steps on FLUX.1-dev (Labs, 2024) and Qwen-Image-20B (Wu et al., 2025) for text-to-image generation and Wan2.1 (Wan et al., 2025) (1.3B and 14B) for text-to-video generation. `sDMD` outperforms the distillation baselines in most metrics on GenEval (Ghosh et al., 2023), DPG (Hu et al., 2024), and WISE (Niu et al., 2025), achieves the highest overall VBench scores on both Wan2.1 models, and improves all reported diversity metrics over the strong DMD2 (Yin et al., 2024a) baseline.

2 Related Work

Diffusion Distillation. To reduce the sampling cost of multi-step diffusion models, trajectory-based distillation methods enable one or few-step generation by learning shortcuts along denoising trajectories (Song et al., 2023; Boffi et al., 2025; Frans et al., 2025). Continuous-time formulations use JVPs (Lu & Song, 2025; Geng et al., 2025), whose efficient computation at scale requires specialized implementations (Zheng et al., 2026b). Recent works aim to avoid JVP computation through specialized architectures that decouple sampling substeps from backbone evaluations (Chen et al., 2026; Shaul et al., 2026). Adversarial approaches use discriminator-based supervision for distillation (Sauer et al., 2023; 2024) with potential training instability (Lin et al., 2025).

Distribution-matching distillation methods align generated and target distributions through f-divergence minimization (Xu et al., 2025; Xie et al., 2024) or score matching (Zhou et al., 2024; Luo et al., 2024). Among these, DMD (Yin et al., 2024b; Luo et al., 2023) minimizes the reverse KL divergence using the difference between teacher and student scores. With their high sample fidelity and scalability, DMD-based methods have been widely adopted for high-resolution image synthesis and video generation (Yin et al., 2025; Huang et al., 2025; Zheng et al., 2026b; Yin et al., 2024a). However, DMD remains sensitive to outliers and the mode-seeking behavior of reverse KL (Bai et al., 2026; Luo et al., 2025). Our work augments DMD with a complementary velocity field to address these limitations and improve sample quality and diversity.

Wasserstein Gradient Flow for Generative Modeling. WGF (Jordan et al., 1998; Ambrosio et al., 2008) evolves a probability distribution by steepest descent of an energy functional under the 2-Wasserstein metric. When a divergence to the target distribution is chosen as the energy, the flow reduces the discrepancy between the two distributions, providing a framework for generative modeling. For example, MMD flows (Arbel et al., 2019) depend on kernel gradients and can provide weak guidance when generated and target samples are far apart relative to the kernel bandwidth. This requires careful kernel design, such as noise-adaptive kernels (Galashov et al., 2025) or kernels built from multiple feature encoders and scales (Feng et al., 2026; Deng et al., 2026).

Sinkhorn divergence provides transport-based guidance (Genevay et al., 2018; Feydy et al., 2019), and recent works (Han et al., 2026; He et al., 2026) use its gradient flow to train one-step generators. However, estimating transport from finite batches can introduce error into the resulting velocity

field (Pooladian & Niles-Weed, 2021; Genevay et al., 2019). Flows based on reverse KL between clean distributions use local score information to guide samples toward the target distribution (Li et al., 2023; Liu, 2017; He et al., 2025). However, these approaches require known target densities or score estimates obtained through methods such as kernel density estimation, which can limit scalability in high-dimensional generation (He et al., 2025; Han et al., 2026). In contrast, DMD offers a scalable diffusion distillation approach by applying reverse KL to noise-perturbed distributions with noise annealing. In this work, we connect DMD to WGF and propose `sDMD`, which incorporates an optimal transport term to improve guidance for outlier samples and promote mode coverage.

3 PRELIMINARIES

Problem Setup and Notation. We work on $\mathbb{R}^d$ and denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability distributions. Let $p \in \mathcal{P}_2(\mathbb{R}^d)$ denote the target distribution and let $p_{\text{ref}} = \mathcal{N}(0, I_d)$ be the reference distribution. We assume access to a pretrained multi-step teacher for p that gives scores at different noise levels. Our goal is to distill this teacher into a student generator $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ that generates a sample in one or a few steps, such that its output distribution $q = (f_\theta)_{\#} p_{\text{ref}}$ approximates p.

Distribution Matching Distillation (DMD). DMD (Yin et al., 2024b) distills a multi-step diffusion teacher into a few-step generator by matching the student distribution q to the teacher distribution p. The original formulation starts from reverse KL between clean distributions, whose gradient involves the difference between the student and target scores. To approximate the gradient, it introduces diffusion annealing, replacing the clean-distribution scores with scores of noisy distributions. For a few-step generated sample $x = f_\theta(z)$ with $z \sim p_{\text{ref}}$, the forward diffusion process gives $x_t = \alpha_t x + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_d)$ and (α_t, σ_t) defines the noise schedule. We denote the corresponding perturbed marginals by q_t and p_t, then the distribution-matching gradient takes the form:

$$\nabla_{\text{DMD}}(\theta) = -\mathbb{E}_{t,z,\epsilon}\left[\alpha_t \left(\nabla \log p_t(x_t) - \nabla \log q_t(x_t)\right) \frac{\partial f_\theta(z)}{\partial \theta}\right]. \tag{1}$$

In practice, the pretrained teacher provides target scores, while an auxiliary diffusion model learns the student score online from generated samples using a score-matching objective (Vincent, 2011).

Wasserstein Gradient Flow (WGF). Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be an energy functional on distributions that satisfies $\arg\min_q \mathcal{F}(q) = p$. WGF (Ambrosio et al., 2008; Jordan et al., 1998) describes the evolution of the probability density $\{q_\tau\}_{\tau \geq 0}$, where τ denotes gradient-flow time. The evolution is characterized by the following partial differential equation (Villani et al., 2009):

$$\partial_\tau q_\tau + \nabla \cdot (q_\tau V_\tau) = 0, \qquad V_\tau(x) = -\nabla \frac{\delta \mathcal{F}}{\delta q}(q_\tau)(x), \tag{2}$$

where the velocity field V_τ is the negative gradient of the first variation of $\mathcal{F}$. Discretizing this flow in time, each iteration displaces every particle $x_\tau \sim q_\tau$ by an Euler step, i.e., $x_{\tau+1} = x_\tau + V_\tau(x_\tau)$. At the equilibrium $q_\tau = p$, the velocity vanishes and each step leaves every particle in place. Writing the particles as generator output $x = f_\theta(z)$, the fixed-point relation gives the following regression objective on generator parameters (Deng et al., 2026):

$$\mathcal{L}(\theta) = \mathbb{E}_{z \sim p_{\text{ref}}} \| f_\theta(z) - \text{sg}(f_\theta(z) + V_\tau(f_\theta(z)))\|^2, \tag{3}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient, and the loss decreases as $V_\tau \to 0$. Note that different energy functionals induce different velocity fields and shape the training dynamics. Therefore, we connect DMD to WGF to explore the design space of energy functionals for distillation and introduce a complementary transport term to address the limitations of DMD.

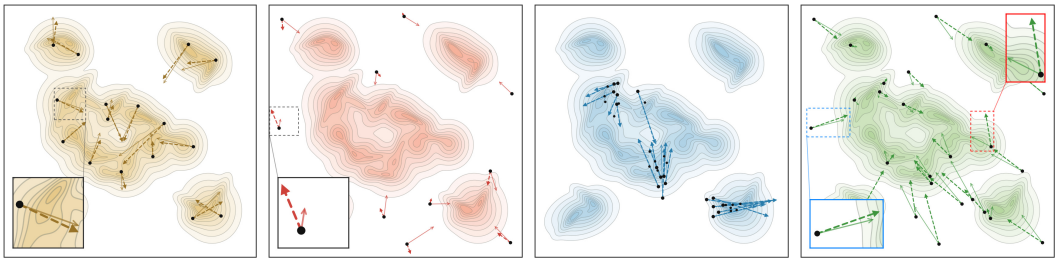
Entropic Optimal Transport and Sinkhorn Divergence. Entropic optimal transport (EOT) compares q and p via a regularized transport cost which can be computed using the Sinkhorn algorithm (Cuturi, 2013; Knight, 2008). For the quadratic cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ and regularization strength $\beta > 0$, EOT is defined as:

$$\text{OT}_\beta(q, p) = \min_{\pi \in \Pi(q,p)} \int c(x, y)\, d\pi(x, y) + \beta D_{\text{KL}}(\pi \,\|\, q \otimes p), \tag{4}$$

where $\Pi(q, p)$ denotes the couplings with marginals q and p. The debiased Sinkhorn divergence removes entropic bias by subtracting the self-transport costs (Genevay et al., 2018):

$$S_\beta(q, p) = \text{OT}_\beta(q, p) - \tfrac{1}{2}\text{OT}_\beta(q, q) - \tfrac{1}{2}\text{OT}_\beta(p, p). \tag{5}$$

Figure 1: **IKL and Sinkhorn Velocity Fields.** We visualize velocity fields from a 2D Gaussian mixture experiment. Black circles mark student particles and contours show target density. Solid (dashed) arrows use exact (learned) teacher scores for IKL and large (small) batches for Sinkhorn. (a) For in-domain student samples, the IKL velocity carries the particles onto the target. (b) Outlier samples under poor initialization, the IKL velocity can be inaccurate in regions of low teacher density. (c) The mode-seeking behavior of IKL can concentrate particles on a subset of the target modes. (d) The Sinkhorn velocity guides particles toward the target modes through the transport plan, remaining robust even for outlier samples (blue box). In practice, finite-sample errors in the Sinkhorn velocity can perturb samples for which IKL already provides reliable guidance (red boxes). We address this using our gated scheduler in Sec. 5.2.



(a) IKL: aligned samples. (b) IKL: outlier samples. (c) IKL: mode seeking. (d) Sinkhorn transport.

S_β is nonnegative and vanishes if and only if $q = p$ (Feydy et al., 2019). Let $\pi_{q,p}$ denote the optimal entropic transport plan in Eq. (4), and define its barycentric projection as $T_{q,p}(x) = \int y \, d\pi_{q,p}(y \mid x)$, the weighted average of target points under the transport plan; $T_{q,q}(x)$ denotes the corresponding projection under the self-transport plan. Then, the WGF velocity can be written as the difference between the two barycentric projections (Genevay et al., 2018):

$$V_{S_\beta}(x) = -\nabla \frac{\delta S_\beta(q,p)}{\delta q}(x) = T_{q,p}(x) - T_{q,q}(x). \tag{6}$$

4 Analyzing DMD through Wasserstein Gradient Flow

4.1 Connection of DMD to WGF

DMD (Yin et al., 2024b) has shown strong empirical performance in few-step generation, but the original derivation of the gradient introduces diffusion noise without a formal connection to WGF. We show that the DMD gradient in Eq. (1) follows from the WGF fixed-point relation in Eq. (3). Specifically, the DMD update can be viewed through the WGF framework by choosing the integral KL (IKL) divergence (Luo et al., 2023), i.e., reverse KL averaged over diffusion noise levels, as the energy functional:

$$\mathcal{F}_{\mathrm{IKL}}(q) = \mathbb{E}_t[D_{\mathrm{KL}}(q_t \,\|\, p_t)], \quad V_{\mathrm{IKL}}(x) \overset{\text{(i)}}{=} \mathbb{E}_{t,\epsilon}[\alpha_t \left(\nabla \log p_t(x_t) - \nabla \log q_t(x_t)\right)], \tag{7}$$

where (i) follows by substituting $\mathcal{F}_{\mathrm{IKL}}$ into the WGF velocity in Eq. (2). Plugging this velocity into the fixed-point relation in Eq. (3) and taking the gradient gives:

$$\nabla_{\mathrm{IKL}}(\theta) \coloneqq \nabla_\theta \mathbb{E}_z\left[\|f_\theta(z) - \mathrm{sg}(f_\theta(z) + V_{\mathrm{IKL}}(f_\theta(z)))\|^2\right] = -2\mathbb{E}_z\left[V_{\mathrm{IKL}}(f_\theta(z)) \frac{\partial f_\theta(z)}{\partial \theta}\right], \tag{8}$$

which recovers the DMD gradient in Eq. (1) up to a constant after substituting V_{IKL} from Eq. (7). We present the detailed derivation in Appendix B.1. This connection allows us to analyze the behavior and limitations of DMD through its velocity field and explore synergistic energies to improve distillation.

Connection to Prior and Concurrent Work. Diff-Instruct (Luo et al., 2023) derives the generator score-difference update directly from IKL, but does not explicitly connect it to WGF. ProlificDreamer (Wang et al., 2023b) derives a WGF update for text-to-3D particles, but does not consider distillation into few-step image or video generators. Concurrent analyses (Gretton et al., 2026; Turan et al., 2026) study WGF fixed-point regression for drifting models (Deng et al., 2026) using kernel-smoothed KL rather than IKL, and also do not explicitly connect DMD to WGF. In this work, we connect DMD to WGF under diffusion-noise-averaged KL from fixed-point regression and use this interpretation to introduce complementary Sinkhorn guidance for diffusion distillation.

4.2 LIMITATIONS OF IKL

DMD (Yin et al., 2024b) and its follow-up methods Fan et al. (2026); Ge et al. (2025) have demonstrated empirical success in capturing fine details and scaling to high-resolution image generation (Fan et al., 2025; Yin et al., 2024a) and video generation (Gu et al., 2026; Zheng et al., 2026b; Huang et al., 2025). However, the gradient of IKL in DMD guides particles by local density alone, which brings two practical limitations. First, the clean reverse KL gradient flow is ill-defined when p and q are mutually singular (Glaser et al., 2021). Although forward diffusion gives p_t and q_t full support, in practice the learned teacher score may remain unreliable for generated samples with low teacher density, particularly at low noise levels. Second, the mode-seeking behavior of reverse KL can cause the student to cover only a subset of the teacher modes.

We illustrate these limitations using an experiment with a 2D Gaussian mixture (Fig. 1). When the two distributions are well aligned, the predicted score is accurate and the particles converge onto the target (Fig. 1(a)). However, when a generated sample lies in a low-density region of the teacher distribution, the predicted score becomes inaccurate, and the resulting velocity fails to carry the particles toward the target under weak diffusion annealing (Fig. 1(b)). With higher noise levels, the sample can be better supported, but the velocity can ignore the local structure of the target distribution, and the mode-seeking behavior of reverse KL can collapse the particles onto a single mode (Fig. 1(c)).

Convergence with Exact OT. These limitations motivate augmenting IKL with an optimal transport (OT) regularizer. We show that IKL alone does not guarantee convergence from arbitrary initializations, whereas the WGF of exact OT contracts toward the target. Under regularity assumptions, a sufficiently strong exact OT term counteracts the negative curvature of IKL and ensures convergence of the combined flow to the target in Wasserstein distance (Theorem 1). We provide a failure case for IKL and a sufficient condition on the OT regularization strength for convergence in Appendix B.2. In the next section, we introduce a practical transport correction that complements IKL.

5 sDMD: SINKHORN DISTRIBUTION MATCHING DISTILLATION

5.1 COMPLEMENTING IKL WITH SINKHORN DIVERGENCE

Despite the favorable convergence properties of exact OT, it can be computationally prohibitive. Instead, we use Sinkhorn divergence, a computationally tractable, debiased form of entropic OT with more favorable sample complexity (Cuturi, 2013; Genevay et al., 2019). We compare this choice with other energy functionals, including forward KL and MMD, in Sec. 6.3.

IKL and Sinkhorn induce complementary velocity fields that are useful in different regimes. IKL captures fine details through local score guidance in regions well supported by the teacher, whereas Sinkhorn uses a global transport plan to guide particles with weak teacher support toward the target and promote mode coverage (Fig. 1(d)).

Combining the two gives `sDMD-U`, which augments $\mathcal{F}_{\mathrm{IKL}}$ with a Sinkhorn divergence term:

$$\mathcal{F}_{\text{sDMD-U}}(q) = \mathbb{E}_t[D_{\mathrm{KL}}(q_t \,\|\, p_t)] + w\, S_\beta(q, p), \tag{9}$$

where the hyperparameter w weights the transport term and recovers $\mathcal{F}_{\mathrm{IKL}}$ at $w = 0$. Both terms are nonnegative, vanish only at $q = p$, and Eq. (9) remains a valid energy functional for Eq. (2). By linearity of the first variation in the energy, the velocity field of the combined energy yields the weighted sum of the two fields:

$$\begin{aligned} V_{\text{sDMD-U}}(x) &= V_{\mathrm{IKL}}(x) + w\, V_{S_\beta}(x) \\ &= \mathbb{E}_{t,\epsilon}[\alpha_t\,(\nabla \log p_t(x_t) - \nabla \log q_t(x_t))] + w\,(T_{q,p}(x) - T_{q,q}(x))\,. \end{aligned} \tag{10}$$

In Appendix B.3, we show that the velocity field is identifiable, i.e., $V_{\text{sDMD-U}} \equiv 0$ if and only if $q = p$. The score difference is obtained by passing the perturbed student samples x_t through the teacher and student score networks. Additionally, the transport term is computed using batches of student and teacher samples, i.e., $\{x^{(i)}\}_{i=1}^{M_q} \sim q$ and $\{y^{(i)}\}_{i=1}^{M_p} \sim p$, on which the Sinkhorn-Knopp algorithm returns the two projections of Eq. (6). We discuss the space on which the transport is computed in Sec. 5.3, and study the role of M_q and M_p in Sec. 6.3.

Lastly, substituting Eq. (10) into Eq. (3) gives the training objective of `sDMD-U`:

$$\mathcal{L}_{\text{sDMD-U}}(\theta) = \mathbb{E}_{z\sim p_{\mathrm{ref}}} \| f_\theta(z) - \mathrm{sg}\big(f_\theta(z) + V_{\mathrm{IKL}}(x) + w\, V_{S_\beta}(x)\big)\|^2, \qquad x = f_\theta(z). \tag{11}$$

Convergence with Sinkhorn. We note that Sinkhorn does not automatically inherit the convergence guarantees of exact OT. However, under the assumptions in Appendix B.4, a sufficiently large transport weight and a suitable bound on the Sinkhorn approximation error show convergence of the combined IKL and Sinkhorn flow to the target (Theorem 2). Empirically, `sDMD-U` benefits from the complementary guidance of IKL and Sinkhorn. As shown by the blue box in Fig. 1(d), Sinkhorn provides more informative guidance than IKL for outliers.

In practice, however, the transport plan is computed on finite batches of student and teacher samples, which introduces error in the resulting velocity (Genevay et al., 2019; Pooladian & Niles-Weed, 2021). As a result, applying Sinkhorn at every iteration can perturb well-supported samples for which IKL already provides reliable guidance, as highlighted by the red boxes in Fig. 1(d), while adding computational cost. We therefore seek a criterion that adaptively activates the Sinkhorn correction for samples with relatively weak teacher support, for which IKL guidance may be unreliable.

5.2 Gated Sinkhorn Scheduler

In this section, we propose a practical criterion for identifying samples with relatively weak teacher support that can benefit from Sinkhorn transport. During training, the few-step generator produces oversmoothed or distorted samples that are outside the typical regions of the target distribution. We assume that these outlier samples have low density under the perturbed target distribution p_t but higher density under the perturbed (fake) distribution q_t of student outputs, i.e., $q_t(x_t) \gg p_t(x_t)$. [1]

However, estimating these densities directly is challenging, and we instead exploit the ratio $q_t(x_t)/p_t(x_t)$, with larger values indicating relatively weak teacher support. Specifically, we assess this density imbalance using a discriminator $D_\phi : \mathbb{R}^d \times [0,1] \to (0,1)$ trained to distinguish perturbed target samples from perturbed student samples:

$$\mathcal{L}_{\text{adv}}(\phi) = -\mathbb{E}_{t,\epsilon}[\mathbb{E}_{y\sim p} \log D_\phi(y_t, t) + \mathbb{E}_{x\sim q} \log(1 - D_\phi(x_t, t))] \,, \tag{12}$$

where y_t and x_t are perturbed target and generated samples, respectively. The discriminator is then used to define the following gating criterion:

$$\lambda_\phi(x_t, t) = 1 - D_\phi(x_t, t) \approx \frac{q_t(x_t)}{p_t(x_t) + q_t(x_t)} \in (0,1), \tag{13}$$

where the approximation becomes exact for the optimal discriminator $D^\star(x_t, t) = p_t(x_t)/(p_t(x_t) + q_t(x_t))$. A high value of $\lambda_\phi(x_t, t)$ suggests that x_t has lower density under the target distribution than under the student distribution. We assume that these samples are more likely to be out-of-distribution samples, where the learned teacher score and the resulting IKL (DMD) velocity may be unreliable. We therefore use λ as a threshold and activate the Sinkhorn term when $\lambda_\phi(x_t, t) > \lambda$:

$$\begin{aligned} V_{\text{sDMD}}(x) &= V_{\text{IKL}}(x) + w \cdot \mathbb{1}[\lambda_\phi(x_t, t) > \lambda]\, V_{S_\beta}(x), \\ \mathcal{L}_{\text{sDMD}}(\theta, \phi) &= \mathbb{E}_{z\sim p_{\text{ref}}} \| f_\theta(z) - \text{sg}(f_\theta(z) + V_{\text{sDMD}}(x)) \|^2 + \mathcal{L}_{\text{adv}}(\phi, \theta), \end{aligned} \tag{14}$$

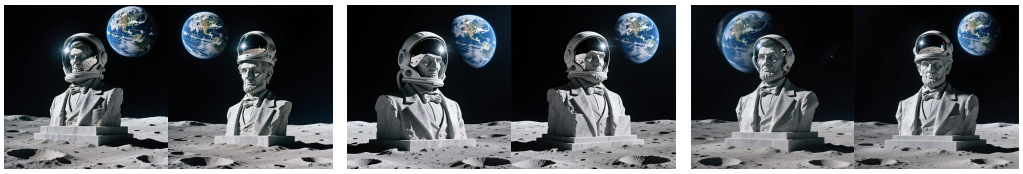
where $\mathbb{1}[\cdot]$ denotes the indicator function. Instead of training a separate discriminator, we incorporate the GAN objective of DMD2 (Yin et al., 2024a) and reuse its discriminator. Here, $\mathcal{L}_{\text{adv}}(\phi, \theta)$ denotes adversarial training with alternating updates to the generator and discriminator. We denote the gated method by `sDMD` and the ungated variant in Eq. (9) by `sDMD-U`, and compare them in Sec. 6.3.

5.3 Practical Considerations

Choice of Transport Space. The transport in Eq. (6) is computed on the sampled latents $x \in \mathbb{R}^d$ of the student and the teacher. Instead, it can be computed in a feature space given a network $\psi : \mathbb{R}^d \to \mathbb{R}^n$, replacing the samples with the feature vectors $\psi(x) \in \mathbb{R}^n$. We empirically find that Sinkhorn transport computed in the clean latent space already provides effective guidance for outlier samples, avoiding the additional cost of a feature encoder. With this choice, `sDMD` incurs only marginal computational overhead compared with DMD2 (Yin et al., 2024a). We present an ablation of the transport space and an analysis of computational cost in Sec. 6.3 and Appendix E, respectively.

[1] This assumption is expected to hold primarily at low to moderate noise levels, as p_t and q_t approach the same prior at high noise levels.

Figure 2: **Qualitative Comparison on Text-to-Image Distillation.** sDMD more closely matches the diversity of the teacher than DMD-based Qwen-Lightning while preserving visual quality. † denotes an officially released checkpoint. See Appendix F for additional results.



Teacher (Qwen-Image) (Wu et al., 2025) Lightning† (Fan et al., 2025) **sDMD** (Ours)

"A towering statue of Abraham Lincoln, cast in a silvery-gray stone, is adorned with a gleaming, opaque astronaut's helmet that reflects the barren lunar landscape. The statue is positioned on the moon's surface, with its craters and dust visible around the base. In the dark sky above, the planet Earth looms large, a swirl of blue and white against the blackness of space."



Teacher (Qwen-Image) (Wu et al., 2025) Lightning† (Fan et al., 2025) **sDMD** (Ours)

"A vibrant and detailed image showcasing a large ceramic bowl filled with steaming ramen, the noodles and broth glistening under the light. Floating amidst the savory dish are several origami boats, each crafted from paper in hues of red, blue, and yellow, adding a whimsical touch to the culinary scene. The bowl sits on a dark wooden table, contrasting with the bright colors of the paper crafts and the rich tones of the ramen ingredients."

Method	NFE	GenEval ↑	DPG ↑	WISE ↑
Teacher	50 × 2	0.860	88.70	0.54
TwinFlow† (Cheng et al., 2026)	2	0.827	86.65	0.50
DMD2 (Yin et al., 2024a)	2	0.856	**87.08**	0.53
sDMD (Ours)	2	**0.862**	87.03	**0.54**

(a) Two-Step Generation.

Method	NFE	GenEval ↑	DPG ↑	WISE ↑
Teacher	50 × 2	0.860	88.70	0.54
Lightning† (Fan et al., 2025)	4	0.864	88.41	0.51
TwinFlow† (Cheng et al., 2026)	4	0.817	86.24	0.49
π-Flow† (Chen et al., 2026)	4	0.863	87.90	0.53
sDMD (Ours)	4	**0.867**	**88.45**	**0.53**

(b) Four-Step Generation.

Table 1: **Qwen-Image Distillation.** Two- and four-step text-to-image generation with Qwen-Image 20B (Wu et al., 2025), evaluated on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and WISE (Niu et al., 2025). † denotes results obtained using officially released checkpoints.

Conditional Generation. For conditional generation, we compute Sinkhorn transport between M_q student samples and M_p teacher samples generated under the same conditioning input. To reduce training cost, we pre-generate the teacher samples for each conditioning input and store them in a buffer. While previous works applied Sinkhorn divergence to class-conditioned generation (Han et al., 2026; He et al., 2026), we extend its use to text-conditioned generation.

6 Experiments

We evaluate sDMD for text-to-image generation with Qwen-Image (Wu et al., 2025) and text-to-video generation with Wan2.1 (Wan et al., 2025). In Appendix, we provide implementation details (Appendix D), additional experimental results (Appendix E), and qualitative results (Appendix F).

6.1 Text-to-Image Diffusion Distillation

We distill Qwen-Image (Wu et al., 2025) (20B) into two- and four-step students at 1024×1024 resolution using teacher samples generated using HPSv2 (Wu et al., 2023) prompts. Following the previous work (Cheng et al., 2026), we evaluate on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and WISE (Niu et al., 2025). We compare our method with state-of-the-art baselines, including TwinFlow (Cheng et al., 2026), π-Flow (Chen et al., 2026), and the DMD2-based Qwen-Lightning (Fan et al., 2025), using their official checkpoints. Two-step DMD2 (Yin et al., 2024a) is trained using FastGen (Nie et al., 2026a) framework.

Results. As shown in Tab. 1, sDMD matches or outperforms the evaluated baselines across all metrics in both two- and four-step generation, except for a marginally lower DPG-Bench score than

Figure 3: **Qualitative Comparison on Text-to-Video Distillation.** We compare sDMD with DMD2 (Yin et al., 2024a) and AnyFlow (Gu et al., 2026) using different seeds, with uniformly sampled frames. † denotes an officially released checkpoint. See Appendix F and the **supplementary webpage** for additional results.



"In this dynamic scene, a surfer expertly maneuvers a wave, showcasing impressive skill and balance. The surfer, clad in a wetsuit, is positioned on a sleek, white surfboard, riding the crest of a powerful, turquoise wave. The wave's size and power are evident from its towering height and the frothy white foam at its peak. The surfer's body leans forward, arms extended for stability, as they navigate the wave's curve with precision. The ocean's deep blue-green hues contrast sharply with the white foam, emphasizing the intensity of the moment. The "Recap" text in the corner suggests this is part of a larger surfing event, highlighting the surfer's performance within a competitive context."



"The video captures an exhilarating bungee jump from a towering cliff. A person, clad in a black outfit with red accents, is seen descending rapidly towards the valley below. The camera follows the jumper's descent, revealing the rugged, rocky cliff face and the expansive green valley dotted with patches of trees. The bungee cord stretches taut, absorbing the impact as the jumper descends. The scene emphasizes the height and scale of the jump, with the vast landscape stretching out beneath the cliff. The dynamic movement of the jumper contrasts sharply with the stillness of the surrounding terrain, highlighting the thrill and adventure of the activity."

Method	NFE	Overall ↑	Quality ↑	Sem. ↑
Teacher	50×2	83.03	84.29	78.00
DMD2 (Yin et al., 2024a)	4	83.65	85.73	75.32
rCM† (Zheng et al., 2026b)	4	82.78	84.13	77.31
AnyFlow† (Gu et al., 2026)	4	83.86	85.15	**78.71**
sDMD (Ours)	4	**84.12**	**85.94**	76.81

(a) Wan2.1 T2V (1.3B).

Method	NFE	Overall ↑	Quality ↑	Sem. ↑
Teacher	50×2	83.71	85.13	78.02
DMD2 (Yin et al., 2024a)	4	84.69	**86.67**	76.75
rCM† (Zheng et al., 2026b)	4	83.97	85.40	78.28
AnyFlow† (Gu et al., 2026)	4	84.38	85.70	79.11
sDMD (Ours)	4	**84.87**	85.96	**80.51**

(b) Wan2.1 T2V (14B).

Table 2: **Wan Text-to-Video Distillation.** We evaluate the generated video samples using VBench (Huang et al., 2024). † denotes results obtained using officially released checkpoints. DMD2 (Yin et al., 2024a) is trained using FastGen (Nie et al., 2026a).

DMD2 in the two-step setting. The qualitative comparisons in Fig. 2 and Appendix F suggest that sDMD more closely matches the diversity of the teacher than the strong Qwen-Lightning baseline (Fan et al., 2025), without degrading visual quality. The quantitative results in Appendix E.3 further show improved diversity over Qwen-Lightning (DMD2) across all reported metrics.

6.2 Text-to-Video Diffusion Distillation

We distill the 1.3B and 14B variants of Wan2.1 (Wan et al., 2025) into four-step students for text-to-video generation. We use teacher samples generated from MovieGenVideoBench prompts (Polyak et al., 2024) for training and evaluate the distilled models on VBench (Huang et al., 2024). We compare sDMD against the official checkpoints of rCM (Zheng et al., 2026b) and AnyFlow (Gu et al., 2026), as well as DMD2 (Yin et al., 2024a) trained on FastGen framework (Nie et al., 2026a).

Results. As shown in Tab. 2b, sDMD achieves the highest overall and semantic scores on Wan2.1-14B, with a quality score comparable to DMD2. Similarly, Tab. 2a shows that sDMD achieves the highest overall and quality scores on Wan2.1-1.3B. The qualitative results in Fig. 3 and the additional examples in Appendix F show that sDMD generates high-quality videos that closely follow the

text prompts with diversity. Consistent with the text-to-image results in Sec. 6.1, the quantitative evaluation in Appendix E.3 shows that sDMD also produces more diverse samples than DMD2 on Wan2.1.

6.3 Ablation Study

Energy Functional. Tab. 3 reports FID on ImageNet-256 for models distilled with different energy functionals. IKL achieves the lowest FID among the individual energy functionals, in line with previous observations that it can help preserve fine details (Zheng et al., 2026b; Huang et al., 2025).

Complementing IKL with Sinkhorn divergence in sDMD further reduces FID. The global transport plan guides samples with weak teacher support toward the target distribution and can alleviate the mode collapse associated with reverse KL, as illustrated in Fig. 1. While MMD can also complement IKL, it yields worse performance than Sinkhorn, consistent with observations in previous work (Han et al., 2026; Gretton et al., 2026; He et al., 2026).

	Energy ($\mathcal{F}$)	FID ↓
Score-based	Forward KL	3.23
	IKL	3.14
Sample-based	MMD	6.21
	Sinkhorn	5.09
Combined	IKL + MMD	3.01
	sDMD (Ours)	**2.83**

Table 3: **Energy Functional.** ImageNet-256 (Deng et al., 2009) distillation with different energy functionals. Combined guidance uses both score- and sample-based updates.

Transport Space and Sinkhorn Scheduler. Tab. 4 evaluates the effects of the transport space and the gated Sinkhorn scheduler on sDMD. Applying Sinkhorn at every iteration (ungated) yields worse performance with more computational overhead. Finite-sample Sinkhorn errors can disrupt reliable IKL guidance (Sec. 5.1), motivating our gated scheduler to activate transport only for samples with weak teacher support.

Space	Scheduler	ImageNet-256	FLUX		
		FID ↓	FID ↓	CMMD ↓	CLIP ↑
$\psi(x)$	Gated	**2.83**	**13.78**	**0.187**	**27.46**
$\psi(x)$	Ungated	2.88	13.89	0.201	**27.46**
x	Gated	2.84	14.48	0.211	27.45

Table 4: **Transport Space and Sinkhorn Scheduler.** Transport in the learned feature space $\psi(x)$ or clean latent space x, with gated (sDMD) or ungated (sDMD−U) Sinkhorn updates.

Computing transport in the learned feature space $\psi(x)$ yields better performance than computing it in the clean latent space x, consistent with observations in previous work (Han et al., 2026; Feng et al., 2026; Deng et al., 2026). However, feature-space transport requires additional forward passes and backpropagation through the feature encoder, whereas clean-latent transport offers competitive quality at a lower computational cost. We therefore compute the Sinkhorn transport in the clean latent space for the large scale generative models (Wan et al., 2025; Wu et al., 2025) to balance sample quality and computational cost.

7 Conclusion

In this work, we proposed sDMD, a diffusion distillation method that combines local score guidance of DMD with global transport. We first connected DMD to WGF by showing that the fixed-point relation with IKL as the energy functional recovers the DMD gradient. This connection enabled us to explore energy functionals that address unreliable guidance for outliers and limited mode coverage in DMD. Specifically, we augmented IKL with Sinkhorn divergence and introduced an adaptive scheduler to focus transport guidance on samples with relatively weak teacher support. Experiments showed that sDMD improves sample quality and diversity over strong diffusion distillation baselines in image and video generation with only marginal training overhead than DMD2.

Limitation and Future Work. For conditional generation, sDMD requires multiple teacher samples for each conditioning input to estimate Sinkhorn guidance. While our ablations show that relatively few samples suffice to provide effective guidance for outlier samples, generating these samples still adds computational costs. Additionally, although our gated Sinkhorn scheduler improves performance, it uses discriminator outputs as a proxy for identifying samples with relatively weak teacher support. Exploring more reliable criteria for identifying samples that benefit from Sinkhorn guidance remains an interesting direction for future work. Lastly, we compute transport in the clean latent space for video diffusion models to reduce computational overhead. Exploring video feature spaces that provide a favorable balance between sample quality and computational cost remains future work.

AI USE STATEMENT

The authors developed the core ideas, including the connection between DMD and Wasserstein gradient flow, our use of Sinkhorn divergence with IKL, and the gated scheduler. These are original ideas developed by the authors, and we did not use AI to generate them. The authors also came up with the core theoretical arguments and proof ideas. We used AI tools to assist with revising and polishing the writings and proofreading the mathematical derviations.

The authors designed the experiments and built the core algorithm and implementation. The authors selected the datasets, evaluation metrics, and baselines. We used AI tools to refactor the codes and implement functionalities of the framework where the authors intervened and verified the outputs.

The authors conducted the related work survey and assessed the relevance of prior work. The authors categorized prior work and established the novelty and originality of our contributions in relation to the existing literature. We did not use AI to conduct the analysis in this survey.

ETHICS STATEMENT

Our method aims to make diffusion generation more efficient and accessible for research and creative applications. However, since our method distills a pretrained model that may have been exposed to harmful content during training, the distilled model may produce harmful or inappropriate content. Additionally, our distillation method does not explicitly address biases in the pretrained model, and the distilled model may inherit these biases.

The authors strongly oppose the use of this method to generate misleading images or videos or to forge content by manipulating identity. Models without adequate safety constraints may nevertheless be vulnerable to such misuse. Stronger safety guardrails during model training and deployment are therefore needed to mitigate these risks.

REPRODUCIBILITY STATEMENT

In Appendix, we provide the algorithm, hardware requirements, and implementation details, including hyperparameters, for all applications. At the time of submission, all datasets used for distillation and pretrained models are publicly available. Our method is implemented using the publicly available FastGen distillation framework, and we plan to release our implementation upon acceptance.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2nd edition, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *NeurIPS*, 2019.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Lichen Bai, Zikai Zhou, Shitong Shao, Wenliang Zhong, Shuo Yang, Shuo Chen, Bojun Chen, and Zeke Xie. Optimizing few-step generation with adaptive matching distillation. *arXiv preprint arXiv:2602.07345*, 2026.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *arXiv preprint arXiv:2505.18825*, 2025.

Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Efficient gradient flows in sliced-wasserstein space. *arXiv preprint arXiv:2110.10972*, 2021.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. In *ICLR*, 2026.

Zhenglin Cheng, Peng Sun, Jianguo Li, and Tao Lin. Twinflow: Realizing one-step generation on large models with self-adversarial flows. In *ICLR*, 2026.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. *arXiv preprint arXiv:2402.05443*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. *arXiv preprint arXiv:2112.02424*, 2021.

Xiangyu Fan, Zesong Qiu, Fanzhou Wang, Peng Gao, and Lei Yang. Qwen-image-lightning, 2025. URL https://github.com/ModelTC/LightX2V-Qwen-Image-Lightning.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *CVPR*, 2026.

Lan Feng, Wuyang Li, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Representation distribution matching for one-step visual generation. *arXiv preprint arXiv:2607.02375*, 2026.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *AISTATS*, 2019.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *ICML*, 2025.

Alexandre Galashov, Valentin De Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *ICLR*, 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. *arXiv preprint arXiv:2506.00523*, 2025.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 1574–1583. PMLR, 2019. URL https://proceedings.mlr.press/v89/genevay19a.html.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *CVPR*, 2026.

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

Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In *NeurIPS*, 2021.

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation. *arXiv preprint arXiv:2605.13724*, 2026.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

Jiajun He, Wenlin Chen, Mingtian Zhang, David Barber, and José Miguel Hernández-Lobato. Training neural samplers with reverse diffusive kl divergence. In *AISTATS*, 2025.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

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

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 1998.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *ICLR*, 2025.

Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 2008.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

Lingxiao Li, Samuel Hurault, and Justin Solomon. Self-consistent velocity matching of probability flows. In *NeurIPS*, 2023.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In *ICML*, 2025.

Dongyang Liu, Gao Peng, David Liu, DU DU, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *ICLR*, 2026.

Qiang Liu. Stein variational gradient descent as gradient flow. In *NeurIPS*, 2017.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *ICLR*, 2025.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

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 *NeurIPS*, 2023.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. In *NeurIPS*, 2024.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *ICCV*, 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. In *ECCV*, 2024.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. In *NeurIPS*, 2021.

Weili Nie, Julius Berner, Chao Liu, and Arash Vahdat. Nvidia fastgen: Fast generation from diffusion models, 2026a. URL https://github.com/NVlabs/FastGen.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition matching distillation for fast video generation. *arXiv preprint arXiv:2601.09881*, 2026b.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Aram-Alexandre Pooladian, Marco Cuturi, and Jonathan Niles-Weed. Debiaser beware: Pitfalls of centering regularized transport maps. In *ICML*, volume 162, pp. 17830–17847. PMLR, 2022. URL https://proceedings.mlr.press/v162/pooladian22a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024.

Neta Shaul, Chao Liu, Arash Vahdat, and Julius Berner. Parallel decoding distillation for fast image and video generation. *arXiv preprint arXiv:2607.26004*, 2026.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *ICML*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

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

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026. URL https://arxiv.org/abs/2603.09936.

Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 2011.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023a.

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. *NeurIPS*, 2023b.

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. In *CVPR*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 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*, 2023.

Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying N Wu, Kevin Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. In *NeurIPS*, volume 37, pp. 45073–45104, 2024.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 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 *ICLR*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *ICLR*, 2026b.

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

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 *ICML*, 2024.

APPENDIX

Supplementary Material. The supplementary material presents a webpage of video comparisons of sDMD with DMD2, rCM, and AnyFlow. To view the comparisons, unzip the supplementary material and open index.html.

A ALGORITHM

Alg. 1 presents sDMD training, with Sinkhorn barycentric projections computed by Alg. 2. We alternate between updating the generator and optimizing the discriminator and student score network. Transport can optionally be computed in the feature space $\psi(x)$ (Sec. 5.3).

Algorithm 1: sDMD training

Input: Teacher score $\nabla \log p_t$, generator f_θ, discriminator D_ϕ, transport weight w, entropic regularization β, gate threshold λ, batch sizes M_q and M_p, Sinkhorn iterations L.

Output: Few-step generator f_θ.

while *not converged* **do**

- Sample $\{z^{(i)}\}_{i=1}^{M_q} \sim p_{\text{ref}}$
- Sample $\{y^{(j)}\}_{j=1}^{M_p} \sim p$
- $\{x^{(i)}\}_{i=1}^{M_q} \leftarrow \{f_\theta(z^{(i)})\}_{i=1}^{M_q}$ // Few-step sampling
- Sample noise levels $\{t_i\}_{i=1}^{M_q}$ and $\{\epsilon^{(i)}\}_{i=1}^{M_q} \sim \mathcal{N}(0, I_d)$
- $x_{t_i}^{(i)} \leftarrow \alpha_{t_i} x^{(i)} + \sigma_{t_i} \epsilon^{(i)}$
- $V_{\text{IKL}}(x^{(i)}) \leftarrow \alpha_{t_i} \left(\nabla \log p_{t_i}(x_{t_i}^{(i)}) - \nabla \log q_{t_i}(x_{t_i}^{(i)}) \right)$ // Eq. (7)
- $V_{\text{sDMD}}(x^{(i)}) \leftarrow V_{\text{IKL}}(x^{(i)})$
- **if** $\lambda_\phi > \lambda$ **then** // Eq. (13)
  - $T_{q,p} \leftarrow \text{Sinkhorn}(\{x^{(i)}\}_{i=1}^{M_q}, \{y^{(j)}\}_{j=1}^{M_p}, \beta, L)$ // Alg. 2
  - $T_{q,q} \leftarrow \text{Sinkhorn}(\{x^{(i)}\}_{i=1}^{M_q}, \{x^{(i)}\}_{i=1}^{M_q}, \beta, L)$
  - $V_{S_\beta}(x^{(i)}) \leftarrow T_{q,p}(x^{(i)}) - T_{q,q}(x^{(i)})$
  - $V_{\text{sDMD}}(x^{(i)}) \leftarrow V_{\text{IKL}}(x^{(i)}) + w\, V_{S_\beta}(x^{(i)})$
- $\mathcal{L}_{\text{sDMD}}(\theta, \phi) \leftarrow \mathbb{E}_{z \sim p_{\text{ref}}} \| f_\theta(z) - \text{sg}(f_\theta(z) + V_{\text{sDMD}}(f_\theta(z)))\|^2 + \mathcal{L}_{\text{adv}}(\phi, \theta)$ // Eq. (14)
- Update (θ, ϕ) with $\nabla_{\theta,\phi} \mathcal{L}_{\text{sDMD}}$
- Update $\nabla \log q_t$ by denoising score matching on $\{x^{(i)}\}_{i=1}^{M_q}$

Algorithm 2: Sinkhorn barycentric projection

Input: Samples $\{x^{(i)}\}_{i=1}^{N}$ and $\{y^{(j)}\}_{j=1}^{M}$, regularization $\beta > 0$, iterations $L \geq 1$.

Output: Barycentric projections $\{T(x^{(i)})\}_{i=1}^{N}$.

Function $\text{Sinkhorn}(\{x^{(i)}\}_{i=1}^{N}, \{y^{(j)}\}_{j=1}^{M}, \beta, L)$**:**

- $a \leftarrow \mathbf{1}_N / N$
- $b \leftarrow \mathbf{1}_M / M$
- $K \leftarrow \left[\exp\left(-\|x^{(i)} - y^{(j)}\|^2 / (2\beta) \right) \right]_{i,j}$
- $v \leftarrow \mathbf{1}_M$
- **for** $\ell = 1, \ldots, L$ **do**
  - $u \leftarrow a \oslash (Kv)$
  - $v \leftarrow b \oslash (K^\top u)$
- $\Pi \leftarrow \text{diag}(u) K \,\text{diag}(v)$
- $T(x^{(i)}) \leftarrow \dfrac{\sum_{j=1}^{M} \Pi_{ij} y^{(j)}}{\sum_{j=1}^{M} \Pi_{ij}}, \quad i = 1, \ldots, N$
- **return** $\{T(x^{(i)})\}_{i=1}^{N}$

B Derivations

B.1 Connection of DMD to WGF

We derive the connection of DMD (Yin et al., 2024b) to WGF presented in Sec. 4.1. Starting from the IKL energy, we derive its WGF velocity and substitute it into the fixed-point relation to recover the DMD gradient.

IKL Energy. Let $q = (f_\theta)_{\#}p_{\text{ref}}$ be the student distribution and p the target distribution. We choose the integral KL divergence (Luo et al., 2023) as the energy functional:

$$\mathcal{F}_{\text{IKL}}(q) = \mathbb{E}_t[D_{\text{KL}}(q_t \,\|\, p_t)]\,, \tag{15}$$

where t denotes the diffusion noise level and the perturbed marginals are:

$$q_t(y) = \int k_t(y \mid x)\, dq(x), \qquad p_t(y) = \int k_t(y \mid x)\, dp(x), \qquad k_t(y \mid x) = \mathcal{N}(y; \alpha_t x, \sigma_t^2 I_d). \tag{16}$$

We use continuous noise schedules on $t \in [0, 1]$ satisfying:

$$\alpha_0 = \sigma_1 = 0, \qquad \alpha_1 = \sigma_0 = 1, \qquad \dot{\alpha}_t > 0, \quad \dot{\sigma}_t < 0 \quad (0 < t < 1). \tag{17}$$

Thus $0 \le \alpha_t, \sigma_t \le 1$, with $p_{t=0} = \mathcal{N}(0, I_d)$ and $p_{t=1} = p$ the clean target. The expectations over t use sampled noise levels in $(0, 1)$, and the noised marginals have positive smooth densities.

IKL Gradient Flow. By Eq. (2), the velocity is the negative spatial gradient of the first variation of $\mathcal{F}_{\text{IKL}}$. To compute this variation, consider a perturbation $q_h = q + h\chi$, where χ is a finite measure with zero total mass and q_h remains in $\mathcal{P}_2(\mathbb{R}^d)$. Writing $\chi_t(y) = \int k_t(y \mid x)\, d\chi(x)$, we have:

$$\begin{aligned} \frac{d}{dh}\mathcal{F}_{\text{IKL}}(q_h)\bigg|_{h=0} &= \mathbb{E}_t\left[\int \chi_t(y)\left(\log\frac{q_t(y)}{p_t(y)} + 1\right) dy\right] \\ &= \int \mathbb{E}_t\left[\int k_t(y \mid x)\left(\log\frac{q_t(y)}{p_t(y)} + 1\right) dy\right] d\chi(x). \end{aligned} \tag{18}$$

Since χ has zero total mass, we can omit the additive constant and write:

$$\frac{\delta \mathcal{F}_{\text{IKL}}}{\delta q}(x) = \mathbb{E}_t\left[\int k_t(y \mid x) \log\frac{q_t(y)}{p_t(y)}\, dy\right]. \tag{19}$$

Using $\nabla_x k_t(y \mid x) = -\alpha_t \nabla_y k_t(y \mid x)$ and integrating by parts yields:

$$\begin{aligned} \nabla_x \frac{\delta \mathcal{F}_{\text{IKL}}}{\delta q}(x) &= -\mathbb{E}_t\left[\alpha_t \int \nabla_y k_t(y \mid x) \log\frac{q_t(y)}{p_t(y)}\, dy\right] \\ &= \mathbb{E}_t\left[\alpha_t \int k_t(y \mid x)\big[\nabla_y \log q_t(y) - \nabla_y \log p_t(y)\big] dy\right]. \end{aligned} \tag{20}$$

Taking the negative gradient and writing the kernel integral gives:

$$\begin{aligned} V_{\text{IKL}}(x) &= -\nabla_x \frac{\delta \mathcal{F}_{\text{IKL}}}{\delta q}(x) \\ &= \mathbb{E}_{t,\epsilon}\big[\alpha_t\big(\nabla \log p_t(x_t) - \nabla \log q_t(x_t)\big)\big]\,, \end{aligned} \tag{21}$$

where $x_t = \alpha_t x + \sigma_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$. This recovers the IKL velocity in Eq. (7).

Recovering the DMD Gradient. Substituting V_{IKL} into Eq. (3) gives:

$$\mathcal{L}(\theta) = \mathbb{E}_{z \sim p_{\text{ref}}} \|f_\theta(z) - \text{sg}(f_\theta(z) + V_{\text{IKL}}(f_\theta(z)))\|^2. \tag{22}$$

The stop-gradient holds the regression target fixed, and only $f_\theta(z)$ contributes to the gradient:

$$\begin{aligned} \nabla_{\text{IKL}}(\theta) \coloneqq \nabla_\theta \mathcal{L}(\theta) &= -2\,\mathbb{E}_z\left[V_{\text{IKL}}(f_\theta(z))\frac{\partial f_\theta(z)}{\partial \theta}\right] \\ &= -2\,\mathbb{E}_{t,z,\epsilon}\left[\alpha_t\big(\nabla \log p_t(x_t) - \nabla \log q_t(x_t)\big)\frac{\partial f_\theta(z)}{\partial \theta}\right] \\ &= 2\,\nabla_{\text{DMD}}(\theta), \end{aligned} \tag{23}$$

where $x_t = \alpha_t f_\theta(z) + \sigma_t \epsilon$. Thus, the fixed-point relation under the IKL energy recovers the DMD gradient in Eq. (1) up to a constant factor, as stated in Eq. (8).

Comparison with Concurrent and Previous Works. DMD (Yin et al., 2024b) uses diffusion-perturbed scores to approximate the clean reverse KL gradient, without an explicit connection to WGF. Diff-Instruct (Luo et al., 2023) derives the IKL gradient for generator training without connecting it to WGF. ProlificDreamer (Wang et al., 2023b) derives a WGF update for 3D particles through a renderer, without considering fixed-point regression. While concurrent works (Gretton et al., 2026; Turan et al., 2026) analyze kernel-smoothed KL and fixed-point training for drifting models, they do not explicitly study IKL averaged over diffusion noise levels or its connection to DMD. We focus on diffusion distillation with pretrained teacher scores, recovering the widely used DMD update through fixed-point regression under the IKL energy. This WGF perspective provides a common framework for deriving and combining velocity fields from complementary energy functionals to address the limitations of DMD in few-step image and video generation.

B.2 IKL Failure and Exact OT Regularization

We first give an IKL failure case, then show how exact OT can restore contraction toward the target.

B.2.1 Stationarity and Negative Curvature of IKL

Take $p = \frac{1}{2}\mathcal{N}(-m, s^2) + \frac{1}{2}\mathcal{N}(m, s^2)$ with $m > s > 0$ and $q_0 = \delta_0$. The perturbed densities are even, their score difference is odd, and the noise at $x = 0$ is symmetric. Consequently $V_{\mathrm{IKL}}(0) = 0$, so $q_\tau = \delta_0$ is a stationary solution of the IKL continuity equation despite $q_\tau \neq p$. This example rules out convergence from arbitrary initializations. Note that it does not show failure from smooth positive densities.

Separately, IKL need not be convex along Wasserstein geodesics, even for a Gaussian target. Take $p = \mathcal{N}(0, b^2)$ and $q_r = \mathcal{N}(0, r^2)$ with $b, r > 0$. Linear interpolation in r is a Wasserstein geodesic, and at each noise level:

$$\begin{aligned} D_{\mathrm{KL}}(q_{r,t} \| p_t) &= \frac{1}{2}\left[\frac{\sigma_t^2 + \alpha_t^2 r^2}{\sigma_t^2 + \alpha_t^2 b^2} - 1 - \log\frac{\sigma_t^2 + \alpha_t^2 r^2}{\sigma_t^2 + \alpha_t^2 b^2}\right], \\ \lim_{r \to 0^+} \frac{d^2}{dr^2} D_{\mathrm{KL}}(q_{r,t} \| p_t) &= -\frac{\alpha_t^4 b^2}{\sigma_t^2(\alpha_t^2 b^2 + \sigma_t^2)} < 0, \end{aligned} \tag{24}$$

where $q_{r,t}$ is the diffusion-perturbed law of q_r. For t sampled from $(0, t_{\max}]$ with $t_{\max} < 1$, the KL curvature remains negative after averaging over t for sufficiently small $r > 0$. While negative curvature alone does not imply nonconvergence, it shows that IKL need not be convex, even for Gaussian targets.

B.2.2 Exact OT Contracts Toward the Target

With our cost convention, exact OT is $\frac{1}{2}W_2^2(q, p)$. For an absolutely continuous source q, let $T_{q,p}^0$ be the optimal transport map to p and define $V_{\mathrm{OT}}(x) = T_{q,p}^0(x) - x$. The superscript 0 denotes zero entropic regularization and $T_{q,p}$ in the main text denotes the entropic barycentric map. The squared-distance chain rule (Ambrosio et al., 2008) gives:

$$\frac{d}{d\tau}\frac{1}{2}W_2^2(q_\tau, p) = -\int V_{\mathrm{OT}}(x)^\top V_\tau(x)\, dq_\tau(x), \tag{25}$$

where V_τ is the flow velocity and V_{OT} is evaluated at the current distribution q_τ. Throughout the convergence arguments, a flow is locally absolutely continuous in W_2, has a density at almost every time, and has square-integrable velocities. Setting $V_\tau = V_{\mathrm{OT}}$ gives:

$$\frac{d}{d\tau}\frac{1}{2}W_2^2(q_\tau, p) = -W_2^2(q_\tau, p), \qquad W_2(q_\tau, p) = e^{-\tau} W_2(q_0, p). \tag{26}$$

B.2.3 Convergence of IKL with Exact OT

Theorem 1 (Convergence of IKL with exact OT). *Let $p \in \mathcal{P}_2(\mathbb{R}^d)$ and consider the energy:*

$$\widetilde{\mathcal{F}}(q) = \mathcal{F}_{\mathrm{IKL}}(q) + \frac{w}{2}W_2^2(q, p), \qquad w > 0. \tag{27}$$

Let q_τ be a Wasserstein gradient flow of Eq. (27)*. Define:*

$$\begin{aligned} M_t &= \sup_y \lambda_{\max}(\mathrm{Cov}_p(X \mid \alpha_t X + \sigma_t \epsilon = y)), \\ L &= \mathbb{E}_t\left[\frac{\alpha_t^4}{\sigma_t^4} M_t\right] < \infty, \qquad X \sim p, \quad \epsilon \sim \mathcal{N}(0, I_d), \end{aligned} \tag{28}$$

where X and ϵ are independent, and $\lambda_{\max}$ denotes the largest eigenvalue of the covariance matrix. Assume differentiation under the integrals and expectations and integration by parts are valid for the OT interpolations in Eq. (30)*. Then the flow satisfies:*

$$W_2(q_\tau, p) \le e^{-(w - L/2)\tau} W_2(q_0, p). \tag{29}$$

When $w > L/2$, the right-hand side of Eq. (29) decays exponentially to zero, proving convergence of the combined IKL and OT flow to p.

Proof. We differentiate IKL along the OT interpolation from the current distribution q to p:

$$\begin{aligned} x_u &= (1-u)x + uT^0_{q,p}(x) = x + uV_{\mathrm{OT}}(x), \\ q_u &= (\mathrm{id} + uV_{\mathrm{OT}})_{\#} q, \qquad q_{u=0} = q, \quad q_{u=1} = p, \quad 0 \le u \le 1. \end{aligned} \tag{30}$$

Here q and its optimal map stay fixed as u varies; τ is flow time and t is the diffusion noise level. Differentiating $q_{u,t}(y) = \int k_t(y \mid x_u)\, dq(x)$ gives:

$$\begin{aligned} \partial_u q_{u,t}(y) &= -\alpha_t \int V_{\mathrm{OT}}(x)^\top \nabla_y k_t(y \mid x_u)\, dq(x), \\ \partial_u^2 q_{u,t}(y) &= \alpha_t^2 \int V_{\mathrm{OT}}(x)^\top \nabla_y^2 k_t(y \mid x_u) V_{\mathrm{OT}}(x)\, dq(x). \end{aligned} \tag{31}$$

Substituting Eq. (31) into the second derivative of KL and integrating by parts yields:

$$\begin{aligned} \frac{d^2}{du^2} D_{\mathrm{KL}}(q_{u,t} \| p_t) &= \int \frac{(\partial_u q_{u,t})^2}{q_{u,t}}\, dy + \int (\partial_u^2 q_{u,t}) \log \frac{q_{u,t}}{p_t}\, dy \\ &= \int \frac{(\partial_u q_{u,t})^2}{q_{u,t}}\, dy \\ &\quad + \alpha_t^2 \int\int k_t(y \mid x_u) V_{\mathrm{OT}}(x)^\top \nabla_y^2 \log \frac{q_{u,t}(y)}{p_t(y)} V_{\mathrm{OT}}(x)\, dy\, dq(x). \end{aligned} \tag{32}$$

To bound the negative curvature of IKL, differentiate $p_t(y) = \int k_t(y \mid x)\, dp(x)$:

$$\nabla_y^2 \log p_t(y) = -\frac{I_d}{\sigma_t^2} + \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov}_p(X \mid \alpha_t X + \sigma_t \epsilon = y). \tag{33}$$

Apply Eq. (33) also to q_u, subtract, and use positive semidefiniteness of covariance and Eq. (28):

$$\begin{aligned} \nabla_y^2 \log \frac{q_{u,t}(y)}{p_t(y)} &= \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov}_{q_u}(X \mid \alpha_t X + \sigma_t \epsilon = y) \\ &\quad - \frac{\alpha_t^2}{\sigma_t^4} \mathrm{Cov}_p(X \mid \alpha_t X + \sigma_t \epsilon = y) \\ &\succeq -\frac{\alpha_t^2}{\sigma_t^4} M_t I_d. \end{aligned} \tag{34}$$

Substitute Eq. (34) into Eq. (32), drop the nonnegative first term, and use $\int k_t(y \mid x_u)\, dy = 1$:

$$\begin{aligned} \frac{d^2}{du^2} \mathcal{F}_{\mathrm{IKL}}(q_u) &= \mathbb{E}_t\left[\frac{d^2}{du^2} D_{\mathrm{KL}}(q_{u,t} \| p_t)\right] \\ &\ge -\mathbb{E}_t\left[\frac{\alpha_t^4}{\sigma_t^4} M_t\right] \int \|V_{\mathrm{OT}}(x)\|^2\, dq(x) \\ &= -L W_2^2(q, p). \end{aligned} \tag{35}$$

To prove convergence, we must control how much the IKL velocity can oppose transport toward p. Indeed, substituting $V_\tau = V_{\mathrm{IKL}} + wV_{\mathrm{OT}}$ into Eq. (25) gives:

$$\frac{d}{d\tau}\frac{1}{2}W_2^2(q_\tau, p) = -\int V_{\mathrm{OT}}(x)^\top V_{\mathrm{IKL}}(x)\, dq_\tau(x) - wW_2^2(q_\tau, p). \tag{36}$$

The OT term decreases the distance, while the IKL term can increase it if the velocity inner product is negative. We use the curvature bound to lower-bound this inner product at an arbitrary current distribution q.

First, Eq. (35) implies that $\mathcal{F}_{\mathrm{IKL}}(q_u) + \frac{L}{2}u^2W_2^2(q, p)$ is convex in u. Therefore, using $q_{u=0} = q$, $q_{u=1} = p$, and $\mathcal{F}_{\mathrm{IKL}}(p) = 0$ gives:

$$\mathcal{F}_{\mathrm{IKL}}(q_u) \le (1-u)\mathcal{F}_{\mathrm{IKL}}(q) + \frac{L}{2}u(1-u)W_2^2(q, p). \tag{37}$$

Subtracting $\mathcal{F}_{\mathrm{IKL}}(q)$, dividing by $u > 0$, and taking $u \to 0^+$ bounds the initial energy increase along the OT direction:

$$\frac{d}{du}\mathcal{F}_{\mathrm{IKL}}(q_u)\Big|_{u=0} \le -\mathcal{F}_{\mathrm{IKL}}(q) + \frac{L}{2}W_2^2(q, p). \tag{38}$$

Next, we express this initial slope as the negative inner product of the two velocities. The interpolation satisfies $\partial_u q_u|_{u=0} = -\nabla_x \cdot (qV_{\mathrm{OT}})$. The first-variation chain rule and integration by parts give:

$$\begin{aligned}
\frac{d}{du}\mathcal{F}_{\mathrm{IKL}}(q_u)\Big|_{u=0} &= \int \frac{\delta \mathcal{F}_{\mathrm{IKL}}}{\delta q}(x)\ \partial_u q_u(x)|_{u=0}\ dx \\
&= -\int \frac{\delta \mathcal{F}_{\mathrm{IKL}}}{\delta q}(x)\nabla_x \cdot \big(q(x)V_{\mathrm{OT}}(x)\big)\, dx \\
&= \int \left[\nabla_x \frac{\delta \mathcal{F}_{\mathrm{IKL}}}{\delta q}(x)\right]^\top V_{\mathrm{OT}}(x)\, dq(x) \\
&= -\int V_{\mathrm{IKL}}(x)^\top V_{\mathrm{OT}}(x)\, dq(x).
\end{aligned} \tag{39}$$

Combining Eqs. (38) and (39) yields the required alignment bound:

$$\int V_{\mathrm{IKL}}(x)^\top V_{\mathrm{OT}}(x)\, dq(x) \ge \mathcal{F}_{\mathrm{IKL}}(q) - \frac{L}{2}W_2^2(q, p). \tag{40}$$

Since $\mathcal{F}_{\mathrm{IKL}}(q) \ge 0$, the inner product cannot fall below $-\frac{L}{2}W_2^2(q, p)$.

Finally, apply this bound at $q = q_\tau$ in Eq. (36):

$$\begin{aligned}
\frac{d}{d\tau}\frac{1}{2}W_2^2(q_\tau, p) &\le -\mathcal{F}_{\mathrm{IKL}}(q_\tau) - (w - L/2)W_2^2(q_\tau, p) \\
&\le -(w - L/2)W_2^2(q_\tau, p).
\end{aligned} \tag{41}$$

Thus $w > L/2$ ensures that OT dominates any opposing IKL contribution. Multiplying by the integrating factor gives

$$\frac{d}{d\tau}\left[e^{2(w-L/2)\tau}W_2^2(q_\tau, p)\right] \le 0.$$

Integrating from 0 to τ and taking square roots yields

$$W_2(q_\tau, p) \le e^{-(w-L/2)\tau}W_2(q_0, p),$$

which proves Eq. (29). □

Corollary 1 (A sufficient weight for Gaussian-mixture targets)**.** *Under the flow and calculus assumptions of Theorem 1, let $p = \sum_j \pi_j \mathcal{N}(\mu_j, s^2 I_d)$ have finitely many components, $s > 0$, and positive weights summing to one. Define:*

$$\Delta = \max_{i,j} \|\mu_i - \mu_j\|, \qquad v_t = \alpha_t^2 s^2 + \sigma_t^2.$$

Then $w > L_{\mathrm{mix}}/2$ suffices for Eq. (29) *with L replaced by:*

$$L_{\mathrm{mix}} = \mathbb{E}_t\left[\frac{\alpha_t^4 s^2}{\sigma_t^2 v_t} + \frac{\alpha_t^4 \Delta^2}{4v_t^2}\right], \tag{42}$$

which is finite when the noise level is bounded away from zero.

Proof. To compute the target posterior covariance, let J be the component index. Gaussian conditioning gives:

$$\begin{aligned}\mathbb{E}[X \mid \alpha_t X + \sigma_t \epsilon = y, J = j] &= \frac{\alpha_t s^2}{v_t} y + \frac{\sigma_t^2}{v_t} \mu_j, \\ \mathrm{Cov}(X \mid \alpha_t X + \sigma_t \epsilon = y, J = j) &= \frac{s^2 \sigma_t^2}{v_t} I_d.\end{aligned} \tag{43}$$

Substituting Eq. (43) into the law of total covariance gives:

$$\begin{aligned}&\mathrm{Cov}(X \mid \alpha_t X + \sigma_t \epsilon = y) \\ &\quad = \frac{s^2 \sigma_t^2}{v_t} I_d + \frac{\sigma_t^4}{v_t^2} \mathrm{Cov}(\mu_J \mid \alpha_t X + \sigma_t \epsilon = y).\end{aligned} \tag{44}$$

To bound L, note that in any unit direction, projected means span an interval of length at most Δ, so their variance is at most $\Delta^2/4$. Using this in Eq. (44) and then Eq. (28) gives:

$$\begin{aligned}M_t &\le \frac{s^2 \sigma_t^2}{v_t} + \frac{\sigma_t^4 \Delta^2}{4 v_t^2}, \\ L &\le \mathbb{E}_t \left[\frac{\alpha_t^4 s^2}{\sigma_t^2 v_t} + \frac{\alpha_t^4 \Delta^2}{4 v_t^2} \right] = L_{\mathrm{mix}} < \infty.\end{aligned}$$

Then apply Theorem 1 with $w > L_{\mathrm{mix}}/2 \ge L/2$. □

B.3 Identifiability of sDMD

We establish identifiability of the velocity induced by Eq. (9), i.e., $V_{\text{sDMD-U}}(x) = 0$ for every $x \in \mathbb{R}^d$ implies $p = q$. Let p, q have compact support, and fix $w, \beta > 0$. Assume $\alpha_t, \sigma_t > 0$ and sampled noise levels bounded away from zero.

IKL Identifiability. Since noise averaging preserves convexity and Gaussian smoothing is injective, the identifiability argument for kernel-smoothed scores (Turan et al., 2026) extends to IKL:

$$V_{\mathrm{IKL}} \equiv 0 \quad \Longleftrightarrow \quad \mathcal{F}_{\mathrm{IKL}}(q) = 0 \quad \Longleftrightarrow \quad q = p.$$

Sinkhorn Identifiability. Following (He et al., 2026), the exact Sinkhorn field satisfies:

$$V_{S_\beta} \equiv 0 \quad \Longleftrightarrow \quad S_\beta(q, p) = 0 \quad \Longleftrightarrow \quad q = p.$$

sDMD Identifiability. Since noising is linear and KL is convex, $\mathcal{F}_{\mathrm{IKL}}$ is convex in q. S_β is also convex in q for the quadratic cost (Feydy et al., 2019). Hence $\mathcal{F}_{\text{sDMD-U}} = \mathcal{F}_{\mathrm{IKL}} + w S_\beta$ is convex under mixtures of distributions. For $q_h = (1-h)q + hp = q + h(p-q)$ with $0 < h \le 1$, convexity gives:

$$\mathcal{F}_{\text{sDMD-U}}(q_h) \le (1-h)\mathcal{F}_{\text{sDMD-U}}(q) + h\mathcal{F}_{\text{sDMD-U}}(p).$$

Rearranging and taking $h \to 0^+$, the definition of the first variation yields:

$$\begin{aligned}\mathcal{F}_{\text{sDMD-U}}(p) - \mathcal{F}_{\text{sDMD-U}}(q) &\ge \lim_{h \to 0^+} \frac{\mathcal{F}_{\text{sDMD-U}}(q_h) - \mathcal{F}_{\text{sDMD-U}}(q)}{h} \\ &= \int \frac{\delta \mathcal{F}_{\text{sDMD-U}}}{\delta q}(x) \big[p(x) - q(x)\big]\, dx.\end{aligned}$$

If $V_{\text{sDMD-U}} \equiv 0$, the first variation is spatially constant, and the integral vanishes as $\int [p(x) - q(x)]\, dx = 1 - 1 = 0$. Since both divergences are nonnegative and vanish at $q = p$, we obtain:

$$0 = \mathcal{F}_{\text{sDMD-U}}(p) \ge \mathcal{F}_{\text{sDMD-U}}(q) \ge 0.$$

Since $w > 0$, both $\mathcal{F}_{\mathrm{IKL}}(q)$ and $S_\beta(q, p)$ must vanish, implying $q = p$. Conversely, $q = p$ makes both velocity terms zero, establishing:

$$V_{\text{sDMD-U}} \equiv 0 \quad \Longleftrightarrow \quad q = p.$$

B.4 Convergence with Sinkhorn Regularization

In this section, we extend Appendix B.2 to Sinkhorn flows Eq. (9). We first analyze how Sinkhorn regularization affects convergence (Theorems 2 and 3), then specialize the analysis to Gaussian distributions (Proposition 1).

B.4.1 Convergence of IKL with Sinkhorn

On compact supports, $S_\beta(q,p) \to \frac{1}{2}W_2^2(q,p)$ as $\beta \to 0^+$ (Genevay et al., 2019). To compare velocities, write $\pi^\beta_{q,p}$ for the entropic plan and $T^\beta_{q,p}$ for its barycentric map. Since $V_{S_\beta} - V_{\mathrm{OT}} = (T^\beta_{q,p} - T^0_{q,p}) - (T^\beta_{q,q} - \mathrm{id})$, the triangle and Jensen inequalities gives:

$$\begin{aligned}&\left(\int \|V_{S_\beta}(x) - V_{\mathrm{OT}}(x)\|^2 \, dq(x)\right)^{1/2} \\ &\quad\le \left(\int \|y - T^0_{q,p}(x)\|^2 \, d\pi^\beta_{q,p}(x,y)\right)^{1/2} + \sqrt{2\mathrm{OT}_\beta(q,q)}.\end{aligned} \tag{45}$$

For fixed p, q the Sinkhorn velocity converges to the OT velocity as $\beta \to 0^+$. This supports using Sinkhorn regularization to approximate the OT guidance in Appendix B.2 at small β.

When the Sinkhorn velocity remains sufficiently close to the exact OT velocity along the flow, the following theorem establishes convergence of IKL and Sinkhorn flow to the target distribution.

Theorem 2 (Convergence under relative velocity control)**.** *Let $p \in \mathcal{P}_2(\mathbb{R}^d)$, $w, \beta > 0$, and q_τ be a regular Wasserstein gradient flow of $\mathcal{F}_{\mathrm{IKL}} + wS_\beta(\cdot, p)$. Assume Eq.* (28) *with $L < \infty$ and the assumptions of Theorem 1. If a fixed $0 \le \eta < 1$ satisfies:*

$$\left(\int \|V_{S_\beta}(x) - V_{\mathrm{OT}}(x)\|^2 \, dq_\tau(x)\right)^{1/2} \le \eta W_2(q_\tau, p) \quad \textit{along the flow}, \tag{46}$$

then:

$$W_2(q_\tau, p) \le e^{-[w(1-\eta) - L/2]\tau} W_2(q_0, p). \tag{47}$$

In particular, $w > L/[2(1-\eta)]$ gives exponential convergence to p.

Proof. At $q = q_\tau$, set $e_\beta = V_{S_\beta} - V_{\mathrm{OT}}$. Applying Cauchy–Schwarz inequality and the identity $\int \|V_{\mathrm{OT}}(x)\|^2 \, dq(x) = W_2^2(q,p)$ gives:

$$\begin{aligned}\int V_{\mathrm{OT}}(x)^\top V_{S_\beta}(x) \, dq(x) &= W_2^2(q,p) + \int V_{\mathrm{OT}}(x)^\top e_\beta(x) \, dq(x) \\ &\ge W_2^2(q,p) - W_2(q,p)\left(\int \|e_\beta(x)\|^2 \, dq(x)\right)^{1/2} \\ &\ge (1-\eta) W_2^2(q,p). \quad \text{(by Eq. (46))}\end{aligned} \tag{48}$$

Using Eqs. (25), (40) and (48):

$$\begin{aligned}\frac{d}{d\tau}\frac{1}{2}W_2^2(q_\tau,p) &= -\int V_{\mathrm{OT}}(x)^\top V_{\mathrm{IKL}}(x) \, dq_\tau(x) - w\int V_{\mathrm{OT}}(x)^\top V_{S_\beta}(x) \, dq_\tau(x) \\ &\le -\mathcal{F}_{\mathrm{IKL}}(q_\tau) - [w(1-\eta) - L/2] W_2^2(q_\tau, p) \\ &\le -[w(1-\eta) - L/2] W_2^2(q_\tau, p).\end{aligned} \tag{49}$$

With $a_\eta = w(1-\eta) - L/2$, the integrating factor gives:

$$\begin{aligned}\frac{d}{d\tau}\left[e^{2a_\eta \tau} W_2^2(q_\tau, p)\right] \le 0, \\ W_2^2(q_\tau, p) \le e^{-2a_\eta \tau} W_2^2(q_0, p).\end{aligned}$$

Taking square roots gives Eq. (47). □

Theorem 3 (Distance bound under general transport error)**.** *Let $p \in \mathcal{P}_2(\mathbb{R}^d)$ and q_τ be a Wasserstein gradient flow with velocity:*

$$V_\tau = V_{\mathrm{IKL}} + w\big[V_{S_{\beta(\tau)}} + \zeta_\tau\big], \tag{50}$$

where $w > 0$, $\beta(\tau) > 0$, and ζ_τ is transport-estimation or solver error. Assume the regularity and calculus conditions of Theorem 1 and the IKL curvature bound Eq. (35)*, with a uniform constant $L < \infty$ along the flow. Let b_τ be a nonnegative, locally integrable error bound satisfying:*

$$\left(\int \|V_{S_{\beta(\tau)}}(x) - V_{\mathrm{OT}}(x) + \zeta_\tau(x)\|^2 \, dq_\tau(x)\right)^{1/2} \le b_\tau. \tag{51}$$

If $w > L/2$, then:

$$W_2(q_\tau, p) \le e^{-(w-L/2)\tau} W_2(q_0, p) + w \int_0^\tau e^{-(w-L/2)(\tau-s)} b_s \, ds. \tag{52}$$

Proof. The Cauchy–Schwarz argument in Eq. (48), with Eq. (51), gives:

$$\int V_{\mathrm{OT}}(x)^\top [V_{S_{\beta(\tau)}}(x) + \zeta_\tau(x)] \, dq_\tau(x) \ge W_2^2(q_\tau, p) - b_\tau W_2(q_\tau, p). \tag{53}$$

Using Eqs. (25), (40) and (53):

$$\begin{aligned} \frac{d}{d\tau}\frac{1}{2} W_2^2(q_\tau, p) &\le -\mathcal{F}_{\mathrm{IKL}}(q_\tau) - (w - L/2) W_2^2(q_\tau, p) + w b_\tau W_2(q_\tau, p) \\ &\le -(w - L/2) W_2^2(q_\tau, p) + w b_\tau W_2(q_\tau, p). \end{aligned} \tag{54}$$

We divide Eq. (54) by $W_2(q_\tau, p)$ where the distance is positive. When the distance is zero, its derivative is zero almost everywhere, and the inequality still holds:

$$\frac{d}{d\tau} W_2(q_\tau, p) \le -(w - L/2) W_2(q_\tau, p) + w b_\tau. \tag{55}$$

Multiplying by $e^{(w-L/2)\tau}$ and integrating proves Eq. (52). □

Corollary 2 (Distance bound under bounded transport error)**.** *Under the assumptions of Theorem 3, suppose $b_\tau \le b$ uniformly for some $b \ge 0$. Then:*

$$\begin{aligned} W_2(q_\tau, p) &\le e^{-(w-L/2)\tau} W_2(q_0, p) + \frac{wb}{w - L/2}\big(1 - e^{-(w-L/2)\tau}\big), \\ \limsup_{\tau \to \infty} W_2(q_\tau, p) &\le \frac{wb}{w - L/2}. \end{aligned} \tag{56}$$

Proof. Substitute $b_s \le b$ into Eq. (52) and integrate the exponential. Taking the limit superior gives the second claim. □

At fixed β, uniform bounds on approximation, estimation, and solver errors give b. The radius $wb/(w - L/2)$ approaches b as $w \to \infty$ for fixed b. Note that this does not mean that the flow cannot converge exactly to p.

B.4.2 Gaussian IKL–Sinkhorn Convergence

For Gaussian targets and initializations, we verify the velocity bound in Eq. (46). This gives exact convergence at fixed, sufficiently small $\beta > 0$.

Proposition 1 (Gaussian IKL–Sinkhorn convergence)**.** *Let $p = \mathcal{N}(m_p, B)$ and $q_0 = \mathcal{N}(m_0, A_0)$ satisfy $0 < \lambda_- I_d \preceq A_0, B \preceq \lambda_+ I_d$. Under the schedule in Eq.* (17)*, assume:*

$$L = \mathbb{E}_t\left[\alpha_t^2 \left(\frac{1}{\sigma_t^2} - \frac{1}{\alpha_t^2 \lambda_+ + \sigma_t^2}\right)\right] < \infty. \tag{57}$$

For $\beta > 0$ and $w \ge 0$, the flow of $\mathcal{F}_{\mathrm{IKL}} + w S_\beta(\cdot, p)$ from q_0 admits a global Gaussian solution $q_\tau = \mathcal{N}(m_\tau, A_\tau)$ with $\lambda_- I_d \preceq A_\tau \preceq \lambda_+ I_d$. Set $\eta = \beta^2 \lambda_+^{3/2}/(8\lambda_-^{7/2})$. If $\eta < 1$ and $w > L/[2(1-\eta)]$, then:

$$W_2(q_\tau, p) \le e^{-[w(1-\eta) - L/2]\tau} W_2(q_0, p). \tag{58}$$

Proof. For $q_\tau = \mathcal{N}(m_\tau, A_\tau)$, set $Z_\tau = A_\tau^{1/2} B A_\tau^{1/2}$ and define the symmetric matrices:

$$C_{\mathrm{IKL},\tau} = \mathbb{E}_t\left[\alpha_t^2 \left\{(\alpha_t^2 A_\tau + \sigma_t^2 I_d)^{-1} - (\alpha_t^2 B + \sigma_t^2 I_d)^{-1}\right\}\right],$$
$$C_{\beta,\tau} = A_\tau^{-1/2}\left[(Z_\tau + \tfrac{\beta^2}{4} I_d)^{1/2} - (A_\tau^2 + \tfrac{\beta^2}{4} I_d)^{1/2}\right] A_\tau^{-1/2}.$$

Gaussian scores in Eq. (21) and the Gaussian maps of (Pooladian et al., 2022) give:

$$V_{\mathrm{IKL}}(x) = -\mathbb{E}_t[\alpha_t^2(\alpha_t^2 B + \sigma_t^2 I_d)^{-1}](m_\tau - m_p) + C_{\mathrm{IKL},\tau}(x - m_\tau),$$
$$V_{S_\beta}(x) = m_p - m_\tau + C_{\beta,\tau}(x - m_\tau). \tag{59}$$

Affine velocities preserve Gaussian distributions. Therefore, using $\mathbb{E}_{q_\tau}[X - m_\tau] = 0$ and $\mathbb{E}_{q_\tau}[(X - m_\tau)(X - m_\tau)^\top] = A_\tau$, the mean and covariance evolve according to:

$$\dot{m}_\tau = -\left(\mathbb{E}_t[\alpha_t^2(\alpha_t^2 B + \sigma_t^2 I_d)^{-1}] + w I_d\right)(m_\tau - m_p),$$
$$\dot{A}_\tau = (C_{\mathrm{IKL},\tau} + w C_{\beta,\tau}) A_\tau + A_\tau (C_{\mathrm{IKL},\tau} + w C_{\beta,\tau}). \tag{60}$$

We show that the covariance remains bounded above and away from zero. For a unit vector u with $A_\tau u = \lambda_- u$, the variance is at its lower limit. Using $B \succeq \lambda_- I_d$ and $Z_\tau \succeq \lambda_- A_\tau$ gives:

$$u^\top C_{\mathrm{IKL},\tau} u = \mathbb{E}_t\left[\alpha_t^2\left(\frac{1}{\alpha_t^2 \lambda_- + \sigma_t^2} - u^\top(\alpha_t^2 B + \sigma_t^2 I_d)^{-1} u\right)\right] \geq 0, \tag{61}$$

$$u^\top C_{\beta,\tau} u = \frac{u^\top (Z_\tau + \frac{\beta^2}{4} I_d)^{1/2} u - \sqrt{\lambda_-^2 + \beta^2/4}}{\lambda_-} \geq 0. \tag{62}$$

Thus $u^\top \dot{A}_\tau u = 2\lambda_- u^\top (C_{\mathrm{IKL},\tau} + w C_{\beta,\tau}) u \geq 0$, and the variance cannot decrease at its lower limit. At the upper limit $A_\tau u = \lambda_+ u$, the inequalities reverse: $u^\top \dot{A}_\tau u \leq 0$, and the variance cannot increase. Hence $\lambda_- I_d \preceq A_\tau \preceq \lambda_+ I_d$ for all τ. These bounds keep the covariance and its inverse bounded.

We next bound the relative velocity error without assuming that A_τ and B commute. The Gaussian OT map gives:

$$V_{\mathrm{OT}}(x) = m_p - m_\tau + A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau) A_\tau^{-1/2}(x - m_\tau). \tag{63}$$

Set $g_\beta(z) = \sqrt{z + \beta^2/4} - \sqrt{z}$. Since $\lambda_-^2 I_d \preceq Z_\tau, A_\tau^2 \preceq \lambda_+^2 I_d$, we use the scalar bound:

$$|g_\beta'(z)| = \frac{\beta^2}{8\sqrt{z}\sqrt{z + \beta^2/4}(\sqrt{z} + \sqrt{z + \beta^2/4})} \leq \frac{\beta^2}{16\lambda_-^3}, \qquad z \in [\lambda_-^2, \lambda_+^2]. \tag{64}$$

Let M, N be symmetric matrices with eigenvalues in $[\lambda_-^2, \lambda_+^2]$. Write their orthonormal eigenvectors as u_i, v_j, with corresponding eigenvalues μ_i, ν_j. Then Eq. (64) implies:

$$\begin{aligned}
\|g_\beta(M) - g_\beta(N)\|_{\mathrm{F}}^2 &= \sum_{i,j} |g_\beta(\mu_i) - g_\beta(\nu_j)|^2 |u_i^\top v_j|^2 \\
&\leq \left(\frac{\beta^2}{16\lambda_-^3}\right)^2 \sum_{i,j} |\mu_i - \nu_j|^2 |u_i^\top v_j|^2 \\
&= \left(\frac{\beta^2}{16\lambda_-^3}\right)^2 \|M - N\|_{\mathrm{F}}^2.
\end{aligned}$$

Apply this with $M = Z_\tau, N = A_\tau^2$ and use $Z_\tau - A_\tau^2 = Z_\tau^{1/2}(Z_\tau^{1/2} - A_\tau) + (Z_\tau^{1/2} - A_\tau)A_\tau$:

$$\begin{aligned}
\|g_\beta(Z_\tau) - g_\beta(A_\tau^2)\|_{\mathrm{F}} &\leq \frac{\beta^2}{16\lambda_-^3} \|Z_\tau - A_\tau^2\|_{\mathrm{F}} \\
&\leq \frac{\beta^2 \lambda_+}{8\lambda_-^3} \|Z_\tau^{1/2} - A_\tau\|_{\mathrm{F}}.
\end{aligned} \tag{65}$$

Let $e_\beta = V_{S_\beta} - V_{\mathrm{OT}}$. The mean terms cancel. Substituting $C_{\beta,\tau}$ and using $(A_\tau^2)^{1/2} = A_\tau$ gives:

$$\begin{aligned} e_\beta(x) &= \Big[C_{\beta,\tau} - A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)A_\tau^{-1/2}\Big]\,(x - m_\tau) \\ &= A_\tau^{-1/2}\Big[(Z_\tau + \tfrac{\beta^2}{4} I_d)^{1/2} - Z_\tau^{1/2} \\ &\qquad - (A_\tau^2 + \tfrac{\beta^2}{4} I_d)^{1/2} + A_\tau\Big] A_\tau^{-1/2}(x - m_\tau) \\ &= A_\tau^{-1/2}[g_\beta(Z_\tau) - g_\beta(A_\tau^2)]A_\tau^{-1/2}(x - m_\tau). \end{aligned}$$

Substitute $X = m_\tau + A_\tau^{1/2}\xi$ with $\xi \sim \mathcal{N}(0, I_d)$. Using $\mathbb{E}\|K\xi\|^2 = \|K\|_{\mathrm{F}}^2$ for any matrix K gives:

$$\left(\int \|e_\beta(x)\|^2 \, dq_\tau(x)\right)^{1/2} = \|A_\tau^{-1/2}[g_\beta(Z_\tau) - g_\beta(A_\tau^2)]\|_{\mathrm{F}}. \tag{66}$$

For the transport distance, evaluate the exact OT velocity at the same X:

$$V_{\mathrm{OT}}(X) = m_p - m_\tau + A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)\xi.$$

The mean cross term vanishes because $\mathbb{E}[\xi] = 0$. Applying the same identity to the remaining random term gives:

$$\begin{aligned} W_2^2(q_\tau, p) &= \mathbb{E}\|V_{\mathrm{OT}}(X)\|^2 \\ &= \|m_\tau - m_p\|^2 + \|A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)\|_{\mathrm{F}}^2. \end{aligned} \tag{67}$$

Inserting $A_\tau^{1/2}A_\tau^{-1/2} = I_d$ gives:

$$\begin{aligned} \|Z_\tau^{1/2} - A_\tau\|_{\mathrm{F}} &\le \|A_\tau^{1/2}\|_{\mathrm{op}}\|A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)\|_{\mathrm{F}} \\ &\le \lambda_+^{1/2}\|A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)\|_{\mathrm{F}}. \end{aligned} \tag{68}$$

Using $\|A_\tau^{-1/2}\|_{\mathrm{op}} \le \lambda_-^{-1/2}$ in the first inequality yields:

$$\begin{aligned} \left(\int \|e_\beta(x)\|^2 \, dq_\tau(x)\right)^{1/2} &\le \frac{\beta^2\lambda_+}{8\lambda_-^{7/2}}\|Z_\tau^{1/2} - A_\tau\|_{\mathrm{F}} && \text{by Eqs. (65) and (66)} \\ &\le \frac{\beta^2\lambda_+^{3/2}}{8\lambda_-^{7/2}}\|A_\tau^{-1/2}(Z_\tau^{1/2} - A_\tau)\|_{\mathrm{F}} && \text{by Eq. (68)} \\ &\le \eta W_2(q_\tau, p) && \text{by Eq. (67),} \end{aligned} \tag{69}$$

where $\eta = \beta^2\lambda_+^{3/2}/(8\lambda_-^{7/2})$.

Finally, Gaussian conditioning and $B \preceq \lambda_+ I_d$ give:

$$\begin{aligned} \mathrm{Cov}_p(X \mid \alpha_t X + \sigma_t \epsilon = y) &= (B^{-1} + \alpha_t^2\sigma_t^{-2} I_d)^{-1} \\ &\preceq \frac{\lambda_+\sigma_t^2}{\alpha_t^2\lambda_+ + \sigma_t^2} I_d. \end{aligned} \tag{70}$$

Thus M_t in Eq. (28) satisfies:

$$\mathbb{E}_t\left[\frac{\alpha_t^4}{\sigma_t^4} M_t\right] \le \mathbb{E}_t\left[\frac{\alpha_t^4\lambda_+}{\sigma_t^2(\alpha_t^2\lambda_+ + \sigma_t^2)}\right] = L < \infty.$$

Applying Theorem 2 with this curvature bound L and the uniform relative error η proves Eq. (58). □

B.5 Analysis of IKL and Sinkhorn Velocity

Sec. 4.2 discusses two limitations of DMD: unreliable guidance for outlier samples and mode collapse. Outlier behavior depends on multiple factors and is difficult to characterize in general, and we present empirical observations in Figs. 1b and 1d.

In this section, we present analysis on mode coverage behavior using the toy example of (Gretton et al., 2026). We compare IKL velocity of DMD with `sDMD` and show how adding Sinkhorn preserves a correction toward an underrepresented mode when IKL guidance vanishes. We assume exact scores and transport plans. The diffusion coefficients α_t, σ_t control the noise level, while $\beta > 0$ independently controls Sinkhorn regularization.

Two-Mode Setup. Consider the following two-mode distributions (Gretton et al., 2026):

$$p = a\delta_{-D} + (1-a)\delta_D, \qquad q = b\delta_{-D} + (1-b)\delta_D, \qquad D > 0, \quad 0 < b < a < 1. \tag{71}$$

The student has excess mass at D and insufficient mass at $-D$. We evaluate each velocity at $x = D$, where a negative velocity points toward the underrepresented mode.

IKL Velocity. For simplicity, we analyze the IKL update for a single draw of (t, ϵ). A similar analysis applies to the expected update. The stochastic IKL update corresponding to Eq. (21) is:

$$V_{\mathrm{IKL}}(x; t, \epsilon) = \alpha_t \big[\nabla \log p_t(x_t) - \nabla \log q_t(x_t)\big], \qquad x_t = \alpha_t x + \sigma_t \epsilon. \tag{72}$$

Using the Gaussian diffusion kernel from Appendix B.1, the perturbed distributions are:

$$\begin{aligned} p_t(y) &= a\, k_t(y \mid -D) + (1-a)\, k_t(y \mid D), \\ q_t(y) &= b\, k_t(y \mid -D) + (1-b)\, k_t(y \mid D). \end{aligned} \tag{73}$$

Differentiating the two mixtures and subtracting their scores gives:

$$\nabla \log p_t(y) - \nabla \log q_t(y) = -\frac{2\alpha_t D(a-b)}{\sigma_t^2} \frac{k_t(y \mid -D) k_t(y \mid D)}{p_t(y) q_t(y)}. \tag{74}$$

At $x = D$, we have $x_t = \alpha_t D + \sigma_t \epsilon$. To simplify the velocity, define the ratio of the left and right Gaussian kernels at this noisy query:

$$\xi_t(\epsilon) \coloneqq \frac{k_t(x_t \mid -D)}{k_t(x_t \mid D)} = \exp\left[-\frac{2\alpha_t^2 D^2}{\sigma_t^2} - \frac{2\alpha_t D}{\sigma_t}\epsilon\right]. \tag{75}$$

Substituting this ratio into the velocity in Eq. (72) gives:

$$V_{\mathrm{IKL}}(D; t, \epsilon) = -\frac{2\alpha_t^2 D(a-b)}{\sigma_t^2} \frac{\xi_t(\epsilon)}{[1 - a + a\xi_t(\epsilon)][1 - b + b\xi_t(\epsilon)]}. \tag{76}$$

Sinkhorn Velocity. Let $m_\beta(a, b) \coloneqq \pi_{q,p}(D, -D)$ be the mass transported from D to $-D$. A transport plan has student modes as rows and target modes as columns, both in the order $(-D, D)$. Its row sums must be $(b, 1-b)$ and column sums $(a, 1-a)$, so any feasible plan has the form:

$$\pi(m) = \begin{pmatrix} a - m & b - a + m \\ m & 1 - b - m \end{pmatrix}, \qquad a - b \le m \le \min(a, 1-b). \tag{77}$$

For $c(x, y) = \frac{1}{2}\|x - y\|^2$, staying at the same mode costs zero and crossing between modes costs $2D^2$. Thus, Eq. (4) reduces to minimizing over the feasible interval above:

$$m_\beta(a, b) = \arg\min_m \left\{2D^2(b - a + 2m) + \beta D_{\mathrm{KL}}\big(\pi(m) \,\|\, q \otimes p\big)\right\}. \tag{78}$$

Differentiating the objective with respect to m and setting the derivative to zero gives the following condition for the optimal mass $m = m_\beta(a, b)$:

$$\begin{aligned} 0 &= 4D^2 + \beta \log \frac{m(b - a + m)}{(a - m)(1 - b - m)}, \\ m(b - a + m) &= e^{-4D^2/\beta}(a - m)(1 - b - m). \end{aligned} \tag{79}$$

At $m = m_\beta(a, b)$, dividing the destination-weighted masses by the source mass $1 - b$ gives:

$$\begin{aligned} T_{q,p}(D) &= \frac{(-D)m + D(1-b-m)}{1-b} = D - \frac{2D}{1-b} m_\beta(a,b), \\ T_{q,q}(D) &= D - \frac{2D}{1-b} m_\beta(b,b). \end{aligned} \tag{80}$$

Here $m_\beta(b, b)$ is the transported mass under self-transport $q \to q$. Subtracting the two averages gives the Sinkhorn velocity:

$$V_{S_\beta}(D) = T_{q,p}(D) - T_{q,q}(D) = -\frac{2D}{1-b}\big[m_\beta(a,b) - m_\beta(b,b)\big], \tag{81}$$

which recovers the unregularized transport velocity $-2D(a-b)/(1-b)$ of Gretton et al. (2026) as $\beta \to 0$.

`sDMD` Velocity Analysis. We examine two cases where a student sample at D receives diminishing IKL guidance toward the underrepresented target mode at $-D$: (i) approaching the clean diffusion timestep ($t \to 1^-$, $\sigma_t/\alpha_t \to 0$) at fixed D, and (ii) increasing mode separation ($D \to \infty$) at fixed t. In both cases, with active Sinkhorn term, the velocity presents as:

$$V_{\mathrm{sDMD}}(D; t, \epsilon) = V_{\mathrm{IKL}}(D; t, \epsilon) + wV_{S_\beta}(D). \tag{82}$$

(i) Low diffusion noise. First, let the diffusion timestep approach the clean endpoint, with $D > 0$ and ϵ fixed. As $\sigma_t/\alpha_t \to 0$ with $\alpha_t, \sigma_t > 0$, the exponential decay of $\xi_t(\epsilon)$ dominates in Eq. (76), and the IKL velocity vanishes:

$$V_{\mathrm{IKL}}(D; t, \epsilon) \longrightarrow 0. \tag{83}$$

On the other hand, for fixed $w, \beta > 0$, the Sinkhorn correction is independent of diffusion noise. Substituting this limit into Eq. (82) gives:

$$V_{\mathrm{sDMD}}(D; t, \epsilon) = V_{\mathrm{IKL}}(D; t, \epsilon) + wV_{S_\beta}(D) \longrightarrow wV_{S_\beta}(D). \tag{84}$$

Hence, the Sinkhorn term in `sDMD` does not vanish and continues to guide the sample toward the underrepresented mode at $-D$.

(ii) Large mode separation. We can see a similar result when the mode separation becomes large. As $D \to \infty$, $\xi_t(\epsilon)$ decays exponentially in D^2, causing the IKL velocity in Eq. (76) to vanish. Meanwhile, Eq. (79) gives $m_\beta(a, b) \to a - b$ and $m_\beta(b, b) \to 0$, yielding:

$$V_{\mathrm{IKL}}(D; t, \epsilon) \longrightarrow 0, \qquad \frac{V_{\mathrm{sDMD}}(D; t, \epsilon)}{2D} \longrightarrow -w\frac{a-b}{1-b}. \tag{85}$$

As mode separation grows, IKL guidance vanishes, while the Sinkhorn term in `sDMD` continues to guide the sample toward $-D$, with magnitude growing proportionally to D. These results motivate augmenting IKL with Sinkhorn in `sDMD` to retain guidance when the IKL correction becomes weak.

Improving DMD with Finite Sinkhorn Batches. These properties make Sinkhorn alone a natural candidate for distillation. Indeed, concurrent works demonstrate that Sinkhorn can train generators from scratch (Han et al., 2026; He et al., 2026). However, our analysis assumes that p and q are known exactly, while training computes Sinkhorn from finite batches. Finite batches introduce estimation errors in the entropic transport cost and the barycentric maps used to compute the velocity (Genevay et al., 2019; Pooladian & Niles-Weed, 2021).

Therefore, small transport batches can yield inaccurate Sinkhorn guidance. In our ablation, Sinkhorn alone underperforms IKL (Tab. 3). Nevertheless, combining Sinkhorn with IKL improves generation quality even with relatively small transport batches (Tab. 9). We interpret this improvement in terms of their complementary roles: despite its own estimation error, Sinkhorn can provide more informative guidance than IKL for outlier samples, while IKL captures fine details in regions well supported by the teacher (Fig. 1d).

C Additional Related Works

Training Few-Step Generators from Scratch. Recent methods train one- or few-step generators directly from data (Song et al., 2023; Frans et al., 2025; Zhou et al., 2025; Boffi et al., 2025), with often using JVPs during training (Lu & Song, 2025; Geng et al., 2025). Further advances improve stability and sample quality through revised objectives and training strategies (Song & Dhariwal, 2024; Geng et al., 2026; Lu et al., 2026; Zhang et al., 2026), and reduce computation costs through finite-difference approximations (Wang et al., 2026). On the other hand, recent methods use mini-batch particle interactions to align generated and target distributions during training (Han et al., 2026; Deng et al., 2026). Our work distills few-step image and video generators from pretrained multi-step diffusion models widely used in real-world applications (Wu et al., 2025; Wan et al., 2025; Labs, 2024).

Wasserstein Gradient Flow for Generative Modeling. WGF evolves probability distributions by steepest descent of an energy functional under the 2-Wasserstein metric (Ambrosio et al., 2008). The Jordan–Kinderlehrer–Otto (JKO) scheme (Jordan et al., 1998) approximates this flow by minimizing the energy functional at each step, with a squared 2-Wasserstein penalty that keeps the updated distribution close to the distribution from the preceding step. Neural implementations (Mokrov et al., 2021; Fan et al., 2021) optimize a transport map at each JKO step which incur quadratic training cost in the number of JKO steps. More efficient formulations have been proposed (Choi et al., 2024; Bonet et al., 2021), but their evaluation does not extend to the scale and dimensionality of modern generative models. In this work, we apply the WGF framework to diffusion distillation, training few-step image and video generators from large pretrained diffusion teachers.

Distribution Matching Distillation. The high sample quality and scalability of DMD have led to broad adoption in high-resolution image synthesis (Yin et al., 2024a), hybrid consistency training for image and video generation (Zheng et al., 2026b), and autoregressive video generation (Yin et al., 2025; Huang et al., 2025; Yang et al., 2025). Recent advances further improve its training and sampling: SenseFlow (Ge et al., 2025) introduces fake-score alignment, while Decoupled DMD (Liu et al., 2026) uses separate noise schedules for CFG augmentation and distribution matching. Other methods improve few-step distillation by matching intermediate distributions along the denoising trajectory (Luo et al., 2025; Fan et al., 2026). Additionally, recent works combine DMD-based and trajectory-based training objectives for few-step generation (Cheng et al., 2026; Nie et al., 2026b; Gu et al., 2026).

Meanwhile, AMD (Bai et al., 2026) shares our concern about unreliable teacher guidance, but uses reward scores as a diagnostic proxy to adapt distillation updates. We derive the DMD update from WGF under the IKL energy, opening a design space of energy functionals for improving distillation. `sDMD` builds on this formulation by adding complementary Sinkhorn transport to improve sample quality and diversity. Additionally, this transport correction may complement the above DMD training strategies which we leave to future work.

D Implementation Details

This section describes the training setup and configuration for each application. Tab. 5 summarizes the corresponding hyperparameters.

In all applications, we generate teacher samples using the default guidance scale and number of sampling steps for each model. For DMD2 (Yin et al., 2024a) and `sDMD`, we perform one generator update for every four discriminator and fake-score updates. We use one H200 node for ImageNet, FLUX, and Wan 1.3B, two nodes for Qwen-Image, and four nodes for Wan 14B.

ImageNet-256. We distill the pretrained SiT-XL/2 teacher (Ma et al., 2024) (675M) into a one-step, class-conditional generator at 256×256 resolution. Training uses ImageNet-1K images (Deng et al., 2009) encoded into $32 \times 32 \times 4$ latents with a pretrained VAE (Rombach et al., 2022). The data loader groups the batch on each GPU by class, and each transport problem uses 8 clean student samples and 8 clean target samples from the same class. We compute Sinkhorn transport using features from a frozen MAE encoder, following the latent feature extraction protocol of Deng et al. (2026). We use two-batch sampling to estimate the self-transport Sinkhorn term.

	ImageNet-256 Appendix E.1	FLUX.1 Appendix E.2	Qwen-Image Sec. 6.1	Wan2.1 Sec. 6.2
Model architecture				
Backbone	SiT-XL/2	FLUX.1-dev	Qwen-Image	1.3B / 14B
Resolution	256×256	512×512	1024×1024	832×480
Latent resolution	$32 \times 32 \times 4$	$64 \times 64 \times 16$	$128 \times 128 \times 16$	104×60
Trainable parameters	Full	Full	LoRA	LoRA
Student / score LoRA rank	—	—	256 / 64	256 / 64
Optimization				
Batch size	256	128	128	48 / 32
Training iterations	100K	5K	3K	1K
Student / score learning rate	3×10^{-5}	10^{-5}	5×10^{-5}	10^{-5}
Discriminator learning rate	2×10^{-6}	10^{-5}	10^{-5}	10^{-5}
Student / score weight decay	0	0.01	0.01	0.01
Discriminator weight decay	0.01	0.01	0.01	0.01
Distillation				
Student NFE	1	4	2 / 4	4
Teacher guidance scale	1.5	3.5	4	5
Adversarial loss weight	0.003	0.03	0.03	0.03
Sinkhorn transport				
Samples per condition $M_q = M_p$	8	8	8	6 / 8
Self-transport $T_{q,q}$	Independent batch	Diagonal mask	Diagonal mask	Diagonal mask
Regularization strength β	0.05	0.05	1	1 / 0.5
Sinkhorn iterations L	40	10	100	50
Sinkhorn weight w	0.5	0.5	2	0.5
Transport representation	MAE	DINOv2	Latent	Latent
Gating schedule				
Gate threshold λ	0.50	0.55	0.55	0.55
Gate warmup iterations	1,000	1,000	1,000	50

Table 5: **Implementation Configurations.** Hyperparameters used in `sDMD`.

FLUX.1. We distill FLUX.1-dev (Labs, 2024) (12B) into a four-step generator at 512×512 resolution using $64 \times 64 \times 16$ latents. The training set contains $8{,}000$ teacher-generated images from $1{,}000$ GenEval prompts (Ghosh et al., 2023), with eight images per prompt. Each GPU processes two prompts per batch, with $M_q = M_p = 8$ samples per prompt, and transport is computed independently for each prompt. For feature transport, we use a frozen DINOv2 ViT-B/14 backbone (Oquab et al., 2023) with inputs at 224×224 resolution and extract activations from blocks 3, 6, 9, and 12 to construct a multiscale representation.

Qwen-Image. We distill Qwen-Image (Wu et al., 2025) (20B) into two- and four-step generators at 1024×1024 resolution with $128 \times 128 \times 16$ latents. To reduce memory use, we train LoRA (Hu et al., 2021) adapters with rank 256 for the student and rank 64 for the fake score network. Teacher images are generated from the HPSv2 benchmark set (Wu et al., 2023). Following Deng et al. (2026), we mask the diagonal in self-transport to exclude matches between a sample and itself.

Wan2.1. We distill Wan2.1 (Wan et al., 2025) (1.3B and 14B) into four-step generators producing 81-frame videos at 16 frames per second and 832×480 using the latents which have a spatial resolution of 104×60. Training uses teacher video latents generated from the prompts in MovieGen-VideoBench (Polyak et al., 2024).

For Wan 14B, the memory constraints limit the per-GPU batch size. Therefore, we adopt cross-node sample gathering to pool samples conditioned on the same prompt across GPUs for Sinkhorn transport computation. We use the four-step sampling schedule $(0.999, 0.937, 0.833, 0.624, 0)$ provided in the FastGen framework (Nie et al., 2026a).

E Additional Experimental Results

E.1 CLASS-CONDITIONAL IMAGE DIFFUSION DISTILLATION

We evaluate one-step class-conditional generation on ImageNet-256, using SiT-XL/2 (Ma et al., 2024). We also distill from a SiT-XL/2 variant pretrained with representation alignment (REPA) (Yu et al., 2024), which accelerates teacher pretraining. The pretraining and distillation costs are summarized in Tab. 10.

The student uses the same backbone as the teacher and generates samples with a single network function evaluation (NFE). Training uses real ImageNet-256 images encoded with the Stable Diffusion VAE (Rombach et al., 2022), with Sinkhorn transport computed between samples of the same class. We evaluate FID (Heusel et al., 2017) on 50,000 generated images and compare against trajectory-based baselines (Lu & Song, 2025; Geng et al., 2025; Zheng et al., 2026b) and the distribution-matching baseline (Yin et al., 2024a).

Method	NFE	FID ↓
SiT‡ (Ma et al., 2024)	250×2	2.06
SiT + REPA‡ (Yu et al., 2024)	250×2	1.42
DMD2 (Yin et al., 2024a)	1	3.14
sCM (Lu & Song, 2025)	1	5.11
rCM (Zheng et al., 2026b)	1	3.51
MeanFlow‡ (Geng et al., 2025)	1	3.43
sDMD (Ours)	1	**2.83**
REPA + π-Flow‡ (Chen et al., 2026)	1	2.85
REPA + **sDMD** (Ours)	1	**2.62**

Table 6: **ImageNet-256 One-Step Distillation.** The first two rows show the teacher results. ‡ denotes reported metrics in previous works.

Results. As shown in Tab. 6, sDMD outperforms the one-step baselines, including DMD2 (Yin et al., 2024a), supporting the benefit of incorporating Sinkhorn transport into distribution matching distillation. Additionally, distilling from the SiT-XL/2+REPA (Yu et al., 2024) further improves generation quality. We compare the computational cost of distillation with that of training a one-step model from scratch in Tab. 10.

E.2 FLUX

We distill the FLUX.1 12B (Labs, 2024) teacher into a four-step student at 512×512 resolution. Training uses $8,000$ teacher-generated images obtained from $1,000$ prompts in the GenEval (Ghosh et al., 2023), with eight images per prompt. Sample quality is assessed using FID (Heusel et al., 2017), CMMD (Jayasumana et al., 2024), and CLIP score (Radford et al., 2021), computed on the same generated samples, with $30,000$ images from MJHQ (Li et al., 2024) as the reference set. For the baselines, we use the FastGen (Nie et al., 2026a) framework and compare against DMD2 (Yin et al., 2024a) and sCM (Lu & Song, 2025), as training instability prevented us from reproducing rCM (Zheng et al., 2026b) in our experiments.

Method	NFE	FID ↓	CMMD ↓	CLIP ↑
Teacher	50×2	11.10	0.19	27.31
DMD2 (Yin et al., 2024a)	4	16.67	0.22	27.14
sCM (Lu & Song, 2025)	4	14.20	0.25	26.17
sDMD (Ours)	4	**13.78**	**0.19**	**27.46**

Table 7: **FLUX Distillation.** Four-step text-to-image generation at 512×512 resolution with FLUX.1 12B (Labs, 2024).

Results. On FLUX, sDMD outperforms the few-step baselines across all metrics and achieves CMMD and CLIP scores comparable to those of the teacher (Tab. 7). The qualitative comparison on FLUX.1 in Appendix F shows that sDMD preserves fine details while generating more diverse samples than DMD2 (Yin et al., 2024a), consistent with the quantitative diversity results in Appendix E.3.

E.3 SAMPLE DIVERSITY.

We compare the diversity of samples generated by DMD2 (Yin et al., 2024a) and sDMD. For images, following DAS (Kim et al., 2025), we report truncated CLIP entropy (TCE) and mean pairwise LPIPS distance (LPIPS MPD). TCE measures the spread of generated images in CLIP feature space, while LPIPS MPD measures the average perceptual difference between image pairs. For videos, following PDD (Shaul et al., 2026), we generate five videos per prompt and compute mean pairwise ℓ_2 distances in the feature spaces of V-JEPA 2 (Assran et al., 2025) and VideoMAE V2 (Wang et al., 2023a).

As shown in Tab. 8, sDMD achieves higher scores than DMD2 on all reported diversity metrics across applications. These results are consistent with the toy example in Fig. 1 and suggest that the global transport of Sinkhorn can alleviate the mode collapse associated with reverse-KL-based distribution matching.

Method	FLUX		Qwen-Image		Wan 14B	
	TCE ↑	LPIPS MPD ↑	TCE ↑	LPIPS MPD ↑	V-JEPA2 ↑	VideoMAE ↑
Teacher	7.25	0.469	5.86	0.397	28.91	18.36
DMD2 (Yin et al., 2024a)	5.93	0.206	4.72	0.161	16.78	13.74
sDMD (Ours)	6.75	0.348	5.43	0.301	20.68	13.85

Table 8: **Diversity Comparison.** Across the applications, sDMD achieves higher diversity than DMD2, suggesting that Sinkhorn global transport helps mitigate mode collapse.

E.4 Sinkhorn Batch Size

As discussed in Sec. 5.1, while Sinkhorn transport provides global guidance toward the teacher distribution, in practice it is computed using a finite number of samples per condition, which may introduce errors into the transport velocity. To analyze the effect of this finite-sample approximation, we show ablation results on ImageNet-256 (Deng et al., 2009) in Tab. 9a and Tab. 9b, which sweep the student sample count M_q with $M_p = 8$ and the teacher sample count M_p with $M_q = 8$, respectively. In both cases, using too few samples degrades generation quality, with a more pronounced degradation when fewer teacher samples (M_p) are used. However, sDMD recovers robust performances even with relatively low number of samples from each distribution ($M_q \geq 8, M_p \geq 8$). These results suggest that Sinkhorn velocities estimated from relatively few samples can still provide more informative guidance than IKL velocities for out-of-distribution samples, thereby improving generation quality.

M_q	M_p	FID ↓
1	8	3.23
4	8	2.98
8	8	2.83
16	8	2.84

(a) Student sweep ($M_p = 8$).

M_q	M_p	FID ↓
8	1	3.97
8	4	3.02
8	8	2.83
8	16	2.85

(b) Teacher sweep ($M_q = 8$).

Table 9: **Ablation Study of Sinkhorn Batch Size.** sDMD achieves robust performance with relatively small batch sizes ($M_q, M_p \geq 8$).

E.5 Training Cost

Method	#Params	Iter.	NFE	GPU Hours	FID ↓
SiT (Ma et al., 2024)	675M	7M	250×2	2,290	2.06
SiT + REPA (Yu et al., 2024)	675M / 87M	4M	250×2	1,220	1.42
W-Flow‡ (Han et al., 2026)	675M / 49M	200K	1	4,639	1.29
DMD2 (Yin et al., 2024a)	675M	100K	1	114	3.14
sDMD (Diag. Mask)	675M / 49M	100K	1	118	2.91
sDMD (Two Batch)	675M / 49M	100K	1	145	2.83
REPA + **sDMD** (Two Batch)	675M / 49M	100K	1	145	2.62

Table 10: **Training and Sampling Cost.** Comparison of teacher pretraining (Ma et al., 2024; Yu et al., 2024), training a one-step generator from scratch (Han et al., 2026), and distillation. GPU hours for distillation exclude teacher pretraining, whose cost is listed separately in the first two rows. ‡ denotes reported metrics in previous works.

Tab. 10 compares the training and sampling costs of sDMD with DMD2 (Yin et al., 2024a) and training a one-step generator from scratch (Han et al., 2026). All experiments in our cost analysis are conducted on NVIDIA H200 GPUs. For sDMD, we consider two choices of computing the self-transport map $T_{q,q}$ in Eq. (6): diagonal masking (Deng et al., 2026) and two-batch (Han et al., 2026). Diagonal masking reuses a single student batch while preventing each sample from matching itself, whereas two-batch estimation computes transport between two independent student batches. With diagonal masking, sDMD improves generation quality over DMD2 while increasing training cost by only around 3.5%. Using two batches further improves FID at additional training cost, consistent with the observations of Han et al. (2026).

Distillation and Training from Scratch. Recent methods achieve high generation quality by training one-step generators from scratch (Han et al., 2026; Deng et al., 2026; Geng et al., 2025). In our analysis, we observe that teacher pretraining followed by distillation still requires fewer GPU

hours than training a model from scratch (Han et al., 2026) (Tab. 10). Additionally, advances in representation learning further accelerate convergence and improve generation quality of pretrained models (Yu et al., 2024; Zheng et al., 2026a), supporting for pretraining followed by distillation as an efficient practice.

F Additional Qualitative Results

Fig. 4 presents a gallery of images generated by sDMD in four steps. Extending the comparisons in Figs. 2 and 3, we present additional text-to-image results with FLUX.1 (Labs, 2024) and Qwen-Image 20B (Wu et al., 2025), along with text-to-video results with Wan2.1-14B (Wan et al., 2025). For each prompt, we show samples generated with different random seeds to illustrate the fidelity and diversity of sDMD.

Supplementary Video Results. The supplementary material includes a webpage (index.html) with a video showcase of sDMD and qualitative comparisons against DMD2 (Yin et al., 2024a), rCM (Zheng et al., 2026b), and AnyFlow (Gu et al., 2026).

Figure 4: **Teaser.** 4-step generated images using sDMD with FLUX.1 and Qwen-Image.



Figure 5: **Additional Qualitative Comparison on FLUX.1.** sDMD more closely matches the diversity of the teacher than DMD2 while preserving visual quality.



"In the corner of a dimly lit room, a rectangular, charcoal-gray computer box exhibits its sharp edges as it rests beside an elegantly draped, cherry-red silk bow tie. The contrast between the electronic's matte finish and the tie's glossy sheen is striking. The computer appears dormant, while the bow tie seems almost ready to be picked up and worn to an event. Behind these objects, the ambient lighting casts soft shadows against the plain, light-colored wall."



"A hefty, red-painted chainsaw rests prominently on the surface of a sturdy, solid oak table. The metal components of the chainsaw catch the soft glow of the early morning sunlight, causing a shimmering effect. The wood's rich, deep grain stands in contrast to the boldness of the chainsaw, and there are wood shavings scattered nearby, hinting at recent activity."

Figure 6: **Additional Qualitative Comparison on FLUX.1 (Continued).** sDMD more closely matches the diversity of the teacher than DMD2 while preserving visual quality.



"A gleaming stainless steel showerhead mounted on a beige tiled wall, with beads of water steadily dripping from its nozzle onto the glossy white porcelain urinal below. The urinal is flanked by grey dividers, ensuring privacy, and the floor beneath is speckled with small, wet splashes reflecting the dim early morning light filtering through a frosted window just out of view. The metallic fixtures of the showerhead are complemented by the chrome handles and plumbing visible under the urinal, creating a cohesive and functional restroom vignette."



"Inside a warm room with a large window showcasing a picturesque winter landscape, three gleaming ruby red necklaces are elegantly laid out on the plush surface of a deep purple velvet jewelry box. The gentle glow from the overhead light accentuates the rich color and intricate design of the necklaces. Just beyond the glass pane, snowflakes can be seen gently falling to coat the ground outside in a blanket of white."

Figure 7: **Additional Qualitative Comparison on FLUX.1 (Continued).** sDMD more closely matches the diversity of the teacher than DMD2 while preserving visual quality.



"a focused woman wielding a heavy sledgehammer, poised to strike an intricately carved ice sculpture of a goose. The sculpture glistens in the light, showcasing its detailed wings and feathers, standing on a pedestal of snow. Around her, shards of ice are scattered across the ground, evidence of her previous strikes."



"A scenic train journey unfolds during the monsoon season in Kerala, with raindrops rhythmically tapping against the windowpane. Inside one of the train's carriages, a plush koala bear toy, donning a small, colorful hat, is propped up against the glass, peering out at the lush landscape. Outside, numerous tall coconut trees sway gently in the rain, their green fronds glistening with moisture, creating a verdant backdrop as the train meanders through the countryside."

Figure 8: **Additional Qualitative Comparison on FLUX.1 (Continued).** sDMD more closely matches the diversity of the teacher than DMD2 while preserving visual quality.

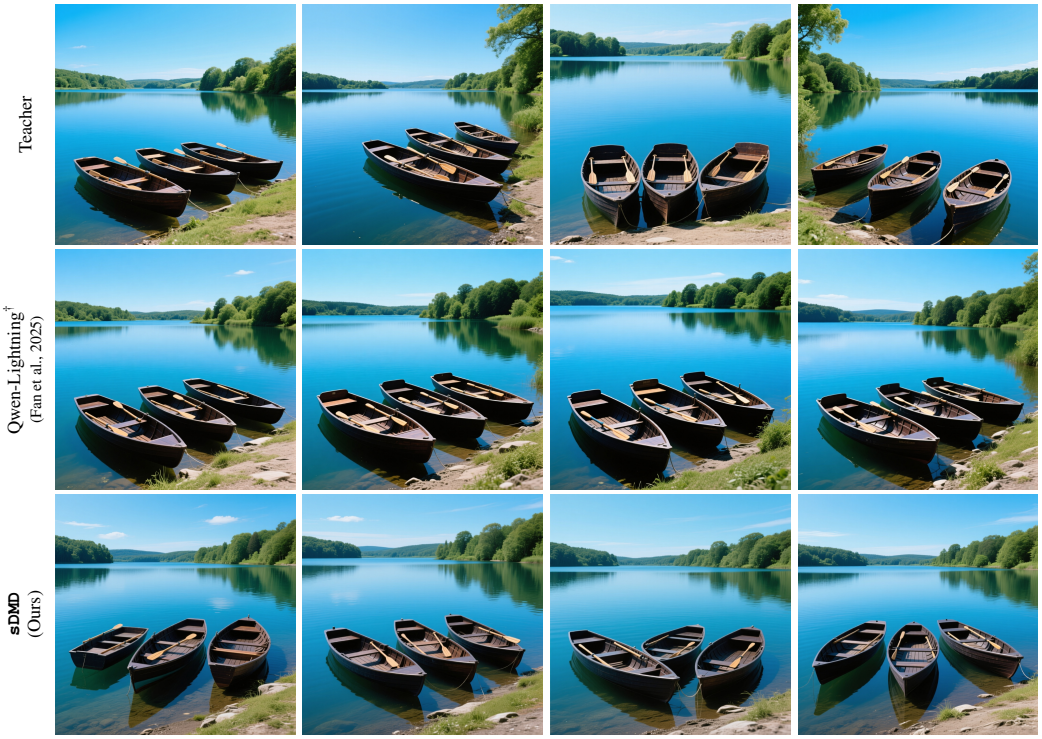


"A whimsical scene where a llama, adorned with a pair of oversized, round sunglasses, stands confidently on the metallic deck of a spacecraft. The deck beneath the llama's hooves gleams with a polished silver finish, reflecting the starry cosmos that surrounds the vessel. In the vast backdrop, the Earth looms large, a swirl of blue oceans and white clouds, providing a stunning contrast to the spaceship's sleek, futuristic design."



"A fluffy white poodle, adorned with a red baseball cap, stands on its hind legs in front of a green chalkboard. In one paw, it holds a large, open dictionary, while the other paw is used to scrawl the word "bonez" in white chalk. The floor beneath the poodle is scattered with colorful pieces of chalk, and the room is filled with small wooden desks and chairs, suggesting a classroom setting."

Figure 9: **Additional Qualitative Comparison on Qwen-Image 20B.** sDMD more closely matches the diversity of the teacher than DMD-based Qwen-Lightning while preserving visual quality. † denotes an officially released checkpoint.



"Three sleek, dark wooden boats are resting along the banks of a tranquil, azure blue lake. The smooth surface of the water reflects the clear sky above and the lush greenery surrounding the lake's edge. Positioned close to each other, the boats' oars are tucked neatly inside, hinting at a recent or upcoming journey across the calm waters."



"An elderly gentleman with a weathered face sits on a park bench, cradling a well-worn wooden guitar in his arms. His fingers are poised over the strings, ready to play a tune. The bench is situated along a path lined with trees that provide a canopy of shade overhead."

Figure 10: **Additional Qualitative Comparison on Qwen-Image 20B (Continued).** sDMD more closely matches the diversity of the teacher than DMD-based Qwen-Lightning while preserving visual quality. † denotes an officially released checkpoint.



"Rows of fresh produce line the interior of a bustling grocery store. Bags filled with crisp apples and plump grapes are neatly arranged on the shelves, inviting shoppers to add them to their carts. The fruits' vibrant colors stand out against the backdrop of other grocery items in the background."

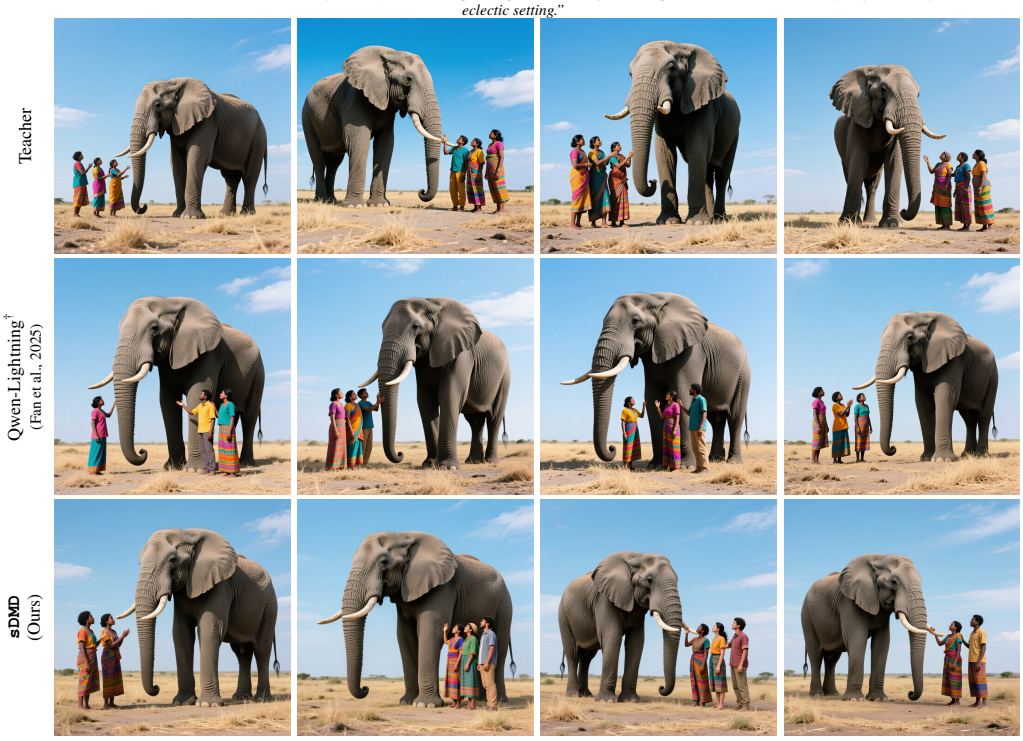


"A scenic trail where a group of riders are mounted on their horses, moving in a line. The horses appear well-groomed, and the riders are dressed in casual riding attire. The path they are on is bordered by trees, and the group seems to be enjoying a leisurely ride in the countryside."

Figure 11: **Additional Qualitative Comparison on Qwen-Image 20B (Continued).** sDMD more closely matches the diversity of the teacher than DMD-based Qwen-Lightning while preserving visual quality. † denotes an officially released checkpoint.



"A stage designed with a fusion of natural and industrial elements, featuring sleek silver glass panels reminiscent of a high-tech laboratory, situated alongside a soft, feathered carpet. The flooring transitions into an area of polished wood leading to a sturdy stone wall, adding an earthy touch to the synthetic ambiance. Overhead, raw concrete merges into the form of a salvia patens roof, juxtaposed with a tranquil sand-laden backdrop that suggests the gentle shores of a lake. In the center, an oval mirror with a slim steel frame reflects the composed space, including the transparent screen that adds a layer of modernity to this eclectic setting."



"A group of three individuals positioned beside a towering elephant in an open area. The elephant's gray skin contrasts with the colorful clothing of the people. They appear to be engaged in a moment of interaction, with the vast sky above them and the ground beneath scattered with dry grass."

Figure 12: **Additional Qualitative Comparison on Wan 14B.** Each method row shows two video samples for the prompt below. † denotes an officially released checkpoint. See the **supplementary webpage** for additional video results.



"The video captures a scene of emergency response following the collapse of the Ji Rao Bridge in Thailand. A group of individuals, including officials and workers, is gathered on the bridge's railing, observing the damage. They appear concerned and engaged in discussion, possibly assessing the situation or coordinating rescue efforts. In the background, heavy machinery, likely excavators, is actively working to clear debris and assess the extent of the damage. The water below flows rapidly, indicating the severity of the flood conditions. The sky is partly cloudy, suggesting variable weather conditions during the incident. This collaborative effort highlights the urgency and scale of the disaster, with multiple parties working together to address the crisis."



"In this video clip, a Korean manicurist is meticulously working on a client's nails. The manicurist's hands, adorned with dark nail polish, are carefully applying a vibrant purple nail polish to the client's fingernails. The client's hands, resting on a pink cushion, are positioned to allow easy access for the manicurist. A bottle of nail polish remover is visible in the background, suggesting a professional setting. The interaction between the manicurist and the client is gentle and precise, highlighting the skill and care involved in the process. The scene captures the intimate and focused nature of a nail treatment session, emphasizing the attention to detail and the personal touch provided by the manicurist."



"The video captures a close-up of a textured white canvas with blue paint being applied using a palette knife. The knife moves methodically across the canvas, spreading the paint in broad, sweeping strokes. The paint appears thick and creamy, creating a raised, almost sculptural effect on the surface. The canvas rests on a dark wooden table, which contrasts sharply with the bright white background of the canvas. The deliberate motion of the knife suggests an artist at work, carefully crafting the texture and form of the painting. The overall scene emphasizes the tactile nature of the painting process, highlighting the interplay between the tool, the paint, and the canvas."

Figure 13: **Additional Qualitative Comparison on Wan 14B (Continued).** Each method row shows two video samples for the prompt below. † denotes an officially released checkpoint. See the **supplementary webpage** for additional video results.



"In the video, a person is seen stirring a pot on a stovetop with a spoon. The pot contains a mixture of what appears to be rice and possibly meat, creating a base for a dish. The individual then adds a generous portion of shredded carrots from a separate bowl into the pot. The carrots are bright orange and appear freshly grated, contrasting with the white rice and brownish hue of the meat. The person uses a large spoon to mix the ingredients thoroughly, ensuring the carrots are evenly distributed throughout the pot. This action suggests the preparation of a hearty, flavorful meal, combining different textures and colors to create a visually appealing and nutritious dish."



"In this dynamic scene from a video game, a player-controlled ship is engaged in a naval battle against two formidable vessels: the Royal Sovereign and HMS Fearless. The player's ship, positioned on the left, maneuvers skillfully amidst turbulent waves, while the enemy ships, larger and more imposing, dominate the right side of the frame. The player's ship is equipped with various controls visible on-screen, including options for "fire swivels," "brace (hold)," "stop," and "speed up." A circular mini-map at the bottom left corner indicates the player's position and nearby objectives. The text overlay provides instructions on targeting weak points, guiding the player through the combat strategy. The overall atmosphere is intense, with the ships battling under a stormy sky, highlighting the strategic and action-packed nature of the gameplay."



"In the video, a person wearing a blue shirt with white polka dots is seen cleaning the trunk of a vehicle. They are using a handheld vacuum cleaner to remove debris from the black carpeted floor, which features a textured pattern. The individual bends down, maneuvering the vacuum to ensure thorough cleaning. The trunk's interior includes a black plastic tray with a grid-like design, likely for organizing items. The person's movements are deliberate and focused, indicating a meticulous approach to maintaining the cleanliness of the vehicle's storage area. The vacuum cleaner is small and portable, suitable for reaching into tight spaces within the trunk. The overall scene suggests a routine maintenance task aimed at keeping the vehicle tidy and functional."