DEMYSTIFYING TRANSITION MATCHING: WHEN AND WHY IT CAN BEAT FLOW MATCHING

Jaihoon Kim[†] **& Minhyuk Sung**
KAIST
{jh27kim,mhsung}@kaist.ac.kr

Rajarshi Saha[‡] **& Youngsuk Park**
Amazon Web Services
{sahrajar,pyoungsu}@amazon.com

ABSTRACT

Flow Matching (FM) underpins many state-of-the-art generative models, yet recent results indicate that Transition Matching (TM) can achieve higher quality with fewer sampling steps. This work answers the question of *when* and *why* TM outperforms FM. First, when the target is a unimodal Gaussian distribution, we prove that TM attains strictly lower KL divergence than FM for finite number of steps. The improvement arises from stochastic difference latent updates in TM, which preserve target covariance that deterministic FM underestimates. We then characterize convergence rates, showing that TM achieves faster convergence than FM under a fixed compute budget, establishing its advantage in the unimodal Gaussian setting. Second, we extend the analysis to Gaussian mixtures and identify local–unimodality regimes in which the sampling dynamics approximate the unimodal case, where TM can outperform FM. The approximation error decreases as the minimal distance between component means increases, highlighting that TM is favored when the modes are well separated. However, when the target variance approaches zero, each TM update converges to the FM update, and the performance advantage of TM diminishes. In summary, we show that TM outperforms FM when the target distribution has well-separated modes and non-negligible variances. We validate our theoretical results with controlled experiments on Gaussian distributions, and extend the comparison to image and video generation.

1 INTRODUCTION

Generative modeling has achieved remarkable success in image and video generation (Chen et al., 2025; Kong et al., 2024; Labs, 2024; Esser et al., 2024; Wan et al., 2025), 3D content creation (Xiang et al., 2025; Li et al., 2025), and scientific applications such as protein and material design (Miller et al., 2025; Geffner et al., 2025). Among these approaches, *Flow Matching* (FM) (Liu et al., 2023; Lipman et al., 2023) has emerged as a core framework and underpins many state-of-the-art generative models (Labs, 2024; Kong et al., 2024). FM learns a continuous-time velocity field and generates samples by numerically simulating the corresponding ordinary differential equation (ODE) with N discretization steps. However, it is computationally demanding because each step requires a full backbone evaluation.

Transition Matching (TM) (Shaul et al., 2025) was recently introduced as an alternative that can surpass FM in the low-step regime. Rather than modeling a velocity field, TM learns a discrete-time transition kernel via a stochastic *difference latent* that specifies per-step updates. At inference, backbone features are cached once per N outer steps, while a lightweight head solves an S-step inner ODE for the difference latent. Although TM converges to FM as $N \to \infty$, it is often superior when N is small, and this advantage still remains largely empirical with limited theoretical understanding of when and why TM outperforms FM.

In this work, we address this gap by providing theoretical analyses clarifying TM's behavior relative to FM. In the analytically tractable unimodal Gaussian setting, we show that for any finite $N > 1$ and $S > 1$, TM achieves strictly lower KL divergence than FM, as TM's *stochastic sampling of the difference latent preserves the target covariance* that FM underestimates. We further establish distinct convergence rates: $\mathcal{O}(1/N^2)$ for FM vs. $\mathcal{O}(1/N^2S^2)$ for TM. Since increasing N requires repeated

[†]Work done during internship at Amazon Web Services.
[‡]Correspondence to Rajarshi Saha.

backbone passes while increasing S only uses a lightweight head, TM consistently attains lower KL divergence under equal compute, explaining its clear advantage in the low-step regime.

We then extend our analysis to mixture of Gaussians, where the target distribution has K components with means $\{\mu_k\}$ and covariances $\{\sigma_k^2 I_d\}$. We show that when the components are *sufficiently well-separated*, sampling effectively reduces to the unimodal case, with approximation error decaying exponentially in the squared minimum separation $D_{\min} = \min_{j \neq k} \|\mu_j - \mu_k\|$. In this regime, TM again outperforms FM at finite steps. However, as variances shrink, the difference latent distribution collapses to its mean, making TM's updates indistinguishable from FM's. This yields a unified view: TM excels when mixture components are *well-separated* and possess *non-negligible variance*.

Finally, we complement our theoretical results with carefully designed experiments that validate the predictions in both unimodal and mixture settings. Beyond synthetic benchmarks, we further evaluate the framework on real-world datasets, including image generation and *for the first time*, video generation. These experiments highlight not only the tight agreement between theory and practice, but also clearly demonstrate that TM consistently outperforms FM under comparable or even lower compute budgets.

Overall, our contributions are summarized as follows:

- **Unimodal Gaussian: Why TM is better.** We prove that TM achieves strictly lower KL divergence than FM for finite $N > 1$ and $S > 1$, can converge faster under fixed compute budget, i.e., $\mathcal{O}(1/S^2)$ vs. $\mathcal{O}(1/S^2N^2)$ (§3; Thm. 1).
- **Gaussian mixtures: When TM is better.** We show that well-separated components reduce to the unimodal case with error decaying exponentially in $D_{\min}$, and TM excels FM when modes are separated and variances are non-negligible (§4; Thm. 2).
- **Validation with real-world applications.** We validate our theoretical insights on two large-scale generative modeling tasks: class-conditioned image generation and frame-conditioned video generation. For images, TM achieves superior quality–compute trade-offs, consistently surpassing FM at a lower latency. For videos, TM consistently outperforms FM upon multiple metrics under matched compute (§5).

2 Overview of Modeling Frameworks

The goal of generative modeling is to construct a process that transports samples from a simple *source distribution* $p_0 = \mathcal{N}(0, I_d)$ to samples from a *target distribution* p_1. This process is formalized as a family of random variables $\{X_t\}_{t\in[0,1]}$, such that $X_0 \sim p_0$ and $X_1 \sim p_1$. In this work, we study Transition Matching (Shaul et al., 2025), a discrete-time, continuous-state framework that generalizes flow matching by learning explicit transition kernels. rather than deterministic mappings.

2.1 Flow Matching (FM)

Flow models (Lipman et al., 2023) describe the continuous evolution of a random variable $\{X_t\}_{t\in[0,1]}$ from an initial distribution $X_0 \sim p_0$ to a target distribution $X_1 \sim p_1$, governed by the ODE,

$$\frac{d}{dt} X_t = u_t(X_t), \qquad X_0 \sim p_0, \tag{1}$$

where $u_t(x) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is a time-dependent *velocity field*. The solution of this ODE is referred to as a flow $\psi_t : \mathbb{R}^d \to \mathbb{R}^d$, with $X_t = \psi_t(X_0)$ evolving deterministically over time. The flow maps the initial distribution along a continuous path of probability densities $\{p_t\}_{t\in[0,1]}$. Therefore, the objective of flow models is to learn the velocity field $u_t(x)$ that induces the desired probability path between source p_0 and target p_1. Flow models use the Conditional Optimal Transport (CondOT) path which linearly interpolates between $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim p_1$:

$$X_t = (1-t)X_0 + tX_1 \sim p_t. \tag{2}$$

To learn the velocity field, flow model parameterizes v_t^θ with a neural network and minimizes the following loss

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1], X_t\sim p_t} \left\| v_t^\theta(X_t) - u_t(X_t) \right\|^2, \tag{3}$$

where $\mathcal{U}[0,1]$ is the uniform distribution. However, $u_t(x)$ in the expression above is generally untractable. Consequently, Conditional Flow Matching (CFM) objective reformulates (3) for the

conditional probability paths $p_{t|1}(x|x_1) = \mathcal{N}(tX_1, (1-t)^2 I_d)$, defined as

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{X_t \sim p_{t|1}, X_1 \sim p_1, t \sim \mathcal{U}[0,1]} \left\| v_t^\theta(X_t) - u_t(X_t|X_1) \right\|^2, \tag{4}$$

where $u_t(X_t \mid X_1) = X_1 - X_0$ is the conditional velocity along the CondOT path (2). Notably, Thm. 2 of Lipman et al. states that minimizing $\mathcal{L}_{\mathrm{CFM}}(\theta)$ and $\mathcal{L}_{\mathrm{FM}}(\theta)$ are equivalent. As the conditional expectation minimizes the squared loss, minimizing $\mathcal{L}_{\mathrm{CFM}}(\theta)$ yields

$$v_t^\theta(X_t) = \mathbb{E}_{X_0, X_1}[X_1 - X_0 | X_t]. \tag{5}$$

Following convention, we refer to the models trained with CFM loss (4) as flow models (FM). At inference time, data is sampled by simulating the ODE in (1) (e.g., Euler method) using the learned v_t^θ. Given a total number of steps N, the interval $[0, 1]$ is discretized as $t_n = n\Delta t$, where $\Delta t = 1/N$ and $n \in [N-1] \triangleq \{0, \ldots, N-1\}$. Each Euler step proceeds as follows:

$$\widetilde{X}_{t_{n+1}}^{\mathrm{FM}} = \widetilde{X}_{t_n}^{\mathrm{FM}} + \Delta t \; v_{t_n}^\theta \left(\widetilde{X}_{t_n}^{\mathrm{FM}} \right), \; n = 0, \ldots, N-1. \tag{6}$$

Here, $\widetilde{X}_0 \sim \mathcal{N}(0, I_d)$, and we use $\widetilde{(\cdot)}$ to distinguish the Euler step sample from the random variables in ODE (1). The superscript FM is used to distinguish the TM sampling steps described next in §2.2.

2.2 Transition Matching (TM)

As outlined in §2.1, FM first learns a continuous time velocity field u_t, which is subsequently discretized to generate a simulation trajectory $\{X_{t_n}\}_{n \in [N-1]}$ using (6). In contrast, TM (Shaul et al., 2025) generalizes this idea to the discrete-time setting by directly learning the probability transition kernel $p\left(X_{t_{n+1}} | X_{t_n}\right)$.

To model the transition kernel $p\left(X_{t_{n+1}} | X_{t_n}\right)$, TM introduces an auxiliary latent variable V,

$$p\left(X_{t_{n+1}} | X_{t_n}\right) = \int p\left(X_{t_{n+1}} | X_{t_n}, V\right) p\left(V | X_{t_n}\right) \mathrm{d}V, \tag{7}$$

where $p\left(X_{t_{n+1}} | X_{t_n}, V\right)$ is chosen to be a deterministic function of (X_{t_n}, V), and the conditional latent distribution $p(V|X_{t_n})$ is learned during training.

To explicitly specify $p\left(X_{t_{n+1}} | X_{t_n}, V\right)$, define $V \triangleq X_1 - X_0$ (henceforth referred to as the *difference latent*), and adopt the linear interpolation (CondOT) path, i.e., $X_{t_n} = (1 - n\Delta t)X_0 + n\Delta t X_1$. Rearranging this, the transition becomes deterministic (given V) as,

$$X_{t_{n+1}} = X_{t_n} + \Delta t V. \tag{8}$$

In other words, TM learns the conditional distribution $p_\theta(V|X_{t_n})$ using a model parametrized by θ, and sampling $V \sim p_\theta(\cdot|X_{t_n})$ drives each discrete transition.

During training, the posterior distribution $p(V|X_{t_n})$ is learned using CFM loss (4). Let $V_1 \triangleq X_1 - X_0$, sample $V_0 \sim \mathcal{N}(0, I_d)$, and choose a linear trajectory, $V_s = (1-s)V_0 + sV_1$, where $s \in [0, 1]$ is used to avoid conflict with t in the FM ODE formulation (1). Then, the TM loss is,

$$\mathcal{L}_{\mathrm{TM}}(\theta) = \mathbb{E}_{\square} \left\| u_s^\theta(V_s | X_{t_n}) - (V_1 - V_0) \right\|^2 \overset{\text{(i)}}{=} \mathbb{E}_{\square} \left\| u_s^\theta(V_s | Z_{t_n}) - (V_1 - V_0) \right\|^2, \tag{9}$$

where, $\square \equiv X_0, V_0 \sim \mathcal{N}(0, I_d)$, $X_1 \sim p_1$, $n \sim \mathcal{U}[N-1]$, $s \sim \mathcal{U}[0, 1]$, and the neural network u_s^θ predicts the velocity field associated with $p(V|X_{t_n})$. Furthermore, with a slight abuse of notation [1], (i) follows from the design of $u_s^\theta(\cdot|\cdot)$, which is composed of a large backbone encoder f_t^θ, that extracts features $Z_{t_n} \triangleq f_t^\theta(X_{t_n})$, followed by a lightweight flow head $u_s^\theta(\cdot|Z_{t_n})$.

Once u_s^θ is trained, for each t_n, a sample from $p(V \mid X_{t_n})$ is obtained by simulating the ODE using $u_s^\theta(\cdot \mid Z_{t_n})$, where $Z_{t_n} = f_t^\theta(X_{t_n})$ remains fixed. Let S be the total number of inner ODE steps, $s \in [S]$ denote the inner step index, and for every n, consider the discretization of $[0, 1]$ as $t_{n,s} = s\Delta s$,where $\Delta s \triangleq 1/S$. Moreover, let $\{\widetilde{X}_{t_n}^{\mathrm{TM}}\}_{n \in [N]}$ denote the samples obtained by simulating the TM

[1] We reuse the notation u_s^θ to denote an output sample of the difference latent V in two contexts: When the input is Z_{t_n} as in $u_s^\theta(\cdot|Z_{t_n})$, it refers to the flow-head, whereas $u_s^\theta(\cdot|X_{t_n})$ refers to flow-head plus the backbone.

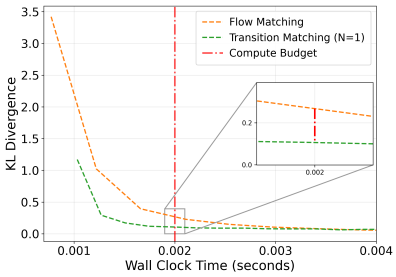

Figure 1: **KL Divergence for Unimodal Gaussian Target.** Comparison of scaling S at $N = 1$ for TM versus scaling N for FM. TM achieves lower KL than FM under a fixed compute.

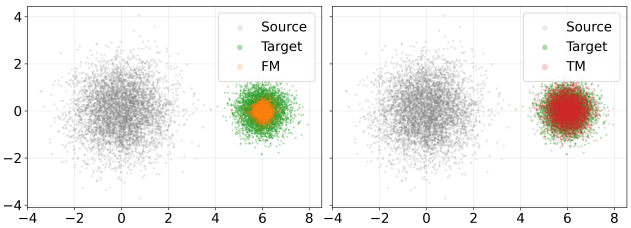

Figure 2: **Qualitative Visualization of Unimodal Gaussian Target.** Each panel shows the source $\mathcal{N}(0, I_d)$ and target $\mathcal{N}(\mu, \sigma^2 I_d)$ with samples from FM (left) and TM (right). With $N = 2$, FM produces reduced variance, whereas TM $(N = 1, S = 2)$ preserves target variance.

dynamics according to (8), and for every n, let $\{\widetilde{V}_{n,s}\}_{s\in[S]}$ be the samples from inner Euler steps. Then, for $n = 0, \ldots, N-1$, and $s = 0, \ldots, S-1$, the TM samples are given by

$$\begin{aligned} \widetilde{X}^{\mathrm{TM}}_{t_{n+1}} &= \widetilde{X}^{\mathrm{TM}}_{t_n} + \Delta t\, \widetilde{V}_{t_n}, \quad \text{where } \widetilde{V}_{t_n} \triangleq \widetilde{V}_{n,1}, \\ \widetilde{V}_{n,s+1} &= \widetilde{V}_{n,s} + \Delta s\, u^\theta_s(\widetilde{V}_{n,s} | Z_{t_n} = f^\theta_t(\widetilde{X}^{\mathrm{TM}}_{t_n})). \end{aligned} \tag{10}$$

Connection to FM. Thm. 1 of Shaul et al. shows that, as the total number of steps $N \to \infty$, the TM update indeed converges to the FM update for every n, that is,

$$\widetilde{X}^{\mathrm{TM}}_{t_{n+1}} \approx \widetilde{X}^{\mathrm{TM}}_{t_n} + \Delta t\, \mathbb{E}\left[X_1 - X_0 | \widetilde{X}^{\mathrm{TM}}_{t_n}\right] \overset{\text{(i)}}{=} \widetilde{X}^{\mathrm{FM}}_{t_{n+1}}, \tag{11}$$

where (i) follows from (5). Since TM and FM coincide in the infinite-step limit, in this work, our focus is instead on the *low-step sampling regime* in the discrete-time setting, where (Shaul et al., 2025) provided only empirical evidence that TM outperformed FM. In the sections that follow, we analyze the mechanisms and conditions that explain *why* and *when* TM can outperform FM.

Scaling Behavior Comparison. We now compare the scaling behavior of FM and TM. As discussed in §2.1, FM requires evaluating the entire backbone network to compute $v^\theta_{t_n}(X_{t_n})$ at each outer step t_n. In contrast, TM computes backbone features once per outer step, i.e., $Z_{t_n} = f^\theta_{t_n}(X_{t_n})$, after which it performs S inner steps with a lightweight flow head $u^\theta_s(\cdot|Z_{t_n})$ to sample a difference latent V. At first glance, it seems that for a fixed number of outer steps N, TM should incur a higher computational cost than FM, since it involves additional inner ODE simulations. However, the backbone computation in TM is amortized across all inner steps, i.e., once Z_{t_n} is obtained, increasing S only requires lightweight flow head evaluations, which are far cheaper than full backbone passes.

Formally, let C_B denote the cost of a single backbone evaluation, (a forward pass either through the network $v^\theta_t(\cdot)$ for FM, or the feature extractor $f^\theta_t(\cdot)$ for TM). Let C_H denote the cost of a single lightweight flow head evaluation $u^\theta_s(\cdot \mid \cdot)$, with $C_H \ll C_B$. The per-sample compute costs are then,

$$\mathcal{C}_{\mathrm{FM}} = NC_B, \qquad \mathcal{C}_{\mathrm{TM}} = NC_B + NSC_H.$$

From this, increasing one outer step in FM incurs a cost of C_B, while the same budget in TM could increase the number of inner steps by

$$\Delta S = \frac{C_B}{NC_H} = \frac{\kappa}{N}, \quad \text{where } \kappa := \frac{C_B}{C_H} \gg 1. \tag{12}$$

This comparison highlights that the computational advantage of TM is pronounced when N is small.

3 Unimodal Gaussian Target

As outlined in the introduction, the unimodal Gaussian target with isotropic covariance provides a simple non-trivial setting in which we explain *why* TM can outperform FM in the low-step sampling regime. Beyond offering formal guarantees, the Gaussian target case also serves as a clean testbed, demonstrating tight agreement between theoretical predictions and empirical results.

Consider $p_0 = \mathcal{N}(0, I_d)$ and $p_1 = \mathcal{N}(\mu, \sigma^2 I_d)$ with the linear path (2), then Thm. (1) below holds.

Theorem 1. *Let $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma > 0$ be independent Gaussian vectors in $\mathbb{R}^d$. If FM and TM iterates follow N Euler steps with $S > 1$ as defined in (6) and (10), respectively, then we have*

$$\mathrm{KL}\left(p_1^{\mathrm{TM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) < \mathrm{KL}\left(p_1^{\mathrm{FM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right),$$

where p_1^{TM} and p_1^{FM} are the marginal distributions of $\widetilde{X}_{t_N=1}^{\mathrm{TM}}$ and $\widetilde{X}_{t_N=1}^{\mathrm{FM}}$, respectively.

The proof is presented in §A.1. The proof shows that both FM and TM exactly match the interpolation mean, so the comparison reduces to covariance. For any finite $N > 1$, the deterministic FM Euler update yields a covariance strictly less than the target covariance which gives a positive KL divergence. In contrast, TM samples a difference latent, and the induced stochasticity effectively compensates the covariance reduction. For any finite $S > 1$, the resulting covariance lies strictly between the FM covariance and the target covariance, so the KL divergence is smaller than that of FM.

Convergence Rate of KL Divergence. Understanding how quickly the KL divergence decays as the number of sampling steps (N) increases provides a concrete quantitative measure of how efficiently p_1^{FM} or p_1^{TM} approach the target distribution p_1, thereby highlighting the tradeoff between computational cost and statistical accuracy. Let $p_1^{\mathrm{FM}}(N)$ and $p_1^{\mathrm{TM}}(N, S)$ denote the marginal distributions of the output iterates of (6) and (10), respectively, with the dependence on the outer and inner Euler steps, N and S, made explicit. The following result precisely characterizes the scaling behavior of both FM and TM samplers.

Corollary 1. *Within the Thm. 3, $p_1^{\mathrm{FM}}(N)$ and $p_1^{\mathrm{TM}}(N, S)$ satisfy*

$$\mathrm{KL}\left(p_1^{\mathrm{FM}}(N) \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \mathcal{O}\left(\frac{1}{N^2}\right), \quad \mathrm{KL}\left(p_1^{\mathrm{TM}}(N, S) \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \mathcal{O}\left(\frac{1}{N^2 S^2}\right). \tag{13}$$

This result follows directly from Thm. 1, with the proof provided in §A.2. Cor. 1 underscores a key distinction: for FM, the KL divergence vanishes only as the number of outer steps N increases, whereas for TM, convergence can instead be achieved by increasing the number of inner ODE steps S, even with N held fixed. As discussed in §2.2, since scaling S in TM is generally more cost-effective than scaling N in FM, this suggests that TM can consistently achieve lower KL divergence under a fixed compute budget.

KL Divergence Across Discretization Steps. To validate the analysis, Fig. 1 reports the KL divergence between the distributions induced by FM and TM for the target $\mathcal{N}(\mu, \sigma^2 I_d)$, illustrating their convergence behavior. The horizontal axis indicates the wall clock time in seconds required to sample the target datapoints. For FM, we vary N, whereas for TM, we fix $N = 1$ and vary only S. Note that since $C_B > C_H$, this setting yields a lower compute cost for TM in the low-step sampling regime (small N).

The empirical curves follow our analysis in Cor. 1. While both samplers converge to zero KL divergence as N and S increase, TM yields lower KL divergence than FM in the low-step sampling regime. Additionally, Fig. 2 visualizes the samples generated by FM ($N = 2$) and TM ($N = 1, S = 2$). We observe that FM produces noticeably smaller sample variance, whereas TM better preserves the target distribution variance even with smaller N, aligning with the quantitative KL divergence trends from Thm. 1.

Analysis of Variance. In the unimodal Gaussian case, the distribution $p(V \mid X_{t_n})$ used to sample difference latent in (10) is itself Gaussian with covariance $\mathrm{Cov}(V \mid X_{t_n}) = \frac{\sigma^2}{(1-t_n)^2 + \sigma^2 {t_n}^2} I_d$, as noted in the proof of Thm. 1 (17). When the target variance $\sigma \to 0$, the covariance of the difference latent distribution correspondingly approaches zero, and the distribution effectively behaves as a delta distribution at its mean. Consequently, each sampled $\widetilde{V}_{t_n}$ of TM becomes nearly identical, concentrating around its expectation $\mathbb{E}[V|X_{t_n}]$. Following (11), each TM update thus reduces to the FM update.

This clearly shows *why* TM outperforms FM in the low-step sampling setting for a unimodal Gaussian target with isotropic covariance $\sigma^2 I_d$ ($\sigma^2 > 0$): *stochastic sampling of the difference latent in TM preserves target variance, whereas deterministic FM updates underestimate it.*

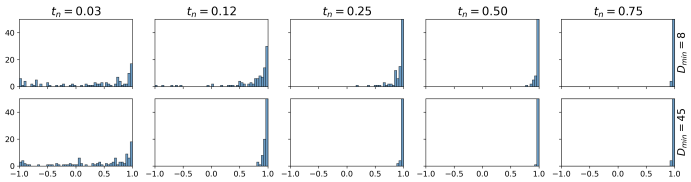

Figure 3: **Effect of $D_{\min}$ on $p(V \mid X)$.** Visualization of $p(V \mid X)$ via cosine-similarity histograms between $\widetilde{V}_{t_n}^{(m)} \sim p(V \mid X_{t_n})$ and $\mathbb{E}[V \mid X_{t_n}]$ for $D_{\min} \in \{8, 45\}$.

4 Mixture of Gaussians Target

Building on the unimodal Gaussian analysis in §3, we now extend the framework to mixture of Gaussians and identify more precisely *when* TM can yield superior performance over FM. Here, we consider the source distribution $p_0 = \mathcal{N}(0, I_d)$, and the target distribution, $p_1 = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2 I_d)$, where $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k > 0$. Unlike the unimodal Gaussian setting, the difference latent distribution $p(V \mid X_t = x)$ is itself a mixture of Gaussian (Bertrand et al., 2025; Karras et al., 2022).

The key observation here is that along parts of the trajectory where a single component dominates (i.e., x lies near the interpolation path mean $t\mu_k$), the mixture distribution of difference latent $p(V|X_t = x)$ closely approximates the corresponding k^{th} component unimodal distribution $p(V|X_t = x, Z = k)$. In other words, this local reduction enables us to leverage the unimodal analysis in §3. Formally, Prop. 1 quantifies the discrepancy of difference latent distribution conditioned on multimodal and unimodal targets.

Proposition 1. *Let $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j^2 I_d)$, where $\sum_j \pi_j = 1,\ \pi_j > 0$, with $X_0 \perp X_1$. Under the linear interpolation path of (2), denote $B_t(j) = (1-t)^2 + t^2\sigma_j^2$ to be the variance of X_t conditioned on $Z = j$, i.e., X_1 being drawn from j^{th} component. Given $x \in \mathbb{R}^d$, let $k_t(x) \in \operatorname{argmin}_j \|x - t\mu_j\|$ (ties resolved arbitrarily), and let $D_t(x) \triangleq \|x - t\mu_{k_t(x)}\|$. Furthermore, define the associated margin,*

$$\rho_t(x) \triangleq \min_{j \neq k_t(x)} \left(\|x - t\mu_j\| - \|x - t\mu_{k_t(x)}\| \right),$$

to denote the gap between the distances to the closest and the second-closest path means. Then,

$$\begin{aligned} &\|p(V|X_t = x) - p(V|X_t = x, Z = k_t(x))\|_{\text{TV}} \\ &\leq C_\pi(x) \left(\frac{B_t^{\max}}{B_t^{\min}} \right)^{\frac{d}{2}} \exp\left(\frac{D_t^2(x)}{2} \left(\frac{1}{B_t^{\min}} - \frac{1}{B_t^{\max}} \right) - \frac{\rho_t^2(x)}{2B_t^{\max}} \right), \end{aligned}$$

where $V = X_1 - X_0$ as before, $C_\pi(x) = \pi_{k_t(x)}^{-1} - 1$, $B_t^{\min} \triangleq \min_j B_t(j)$, and $B_t^{\max} \triangleq \max_j B_t(j)$.

The proof is presented in §A.3. Prop. 1 establishes that the posterior of V given $X_t = x$ can be approximated by the single component conditional $p(V \mid X_t = x, Z = k_t(x))$, which follows the unimodal Gaussian case discussed in §3. The approximation error depends on the mixture weight of the nearest component and the variance $B_t(j)$. Additionally, the bound tightens when x lies near its path mean (small $D_t(x)$) and the closest component is well-separated from others (large $\rho_t(x)$), in which case non-nearest components are exponentially suppressed and the posterior concentrates around a single mode.

We next consider a special case, when the variances are comparable, formalized as $B_t^{\min}/B_t^{\max} \leq 1 - \delta$ for arbitrarily small $\delta \geq 0$, and x lies within a local neighborhood of its nearest path mean, $t\mu_{k_t(x)}$. For clarity of exposition, Cor. 2 considers the case $\delta = 0$.

Corollary 2. *Within the Prop. 1, define the minimal mean separation $D_{\min} \triangleq \min_{j \neq k} \|\mu_j - \mu_k\|$ and the neighborhood $\mathcal{E}_t \triangleq \left\{ x \in \mathbb{R}^d : D_t(x) \leq \sqrt{B_t^{\min}} \right\}$. Assuming $B_t^{\max} = B_t^{\min}$, for any $x \in \mathcal{E}_t$,*

$$\left\| p(V \mid X_t = x) - p(V \mid X_t = x, Z = k_t(x)) \right\|_{\text{TV}} \leq C_\pi(x) \exp\left(2 - \frac{t^2 D_{\min}^2}{4B_t^{\max}} \right).$$

The proof is provided in §A.4, building on Prop. 1. We note that the approximation error decreases with larger mean separation, $D_{\min}$, particularly as $t \to 1$.

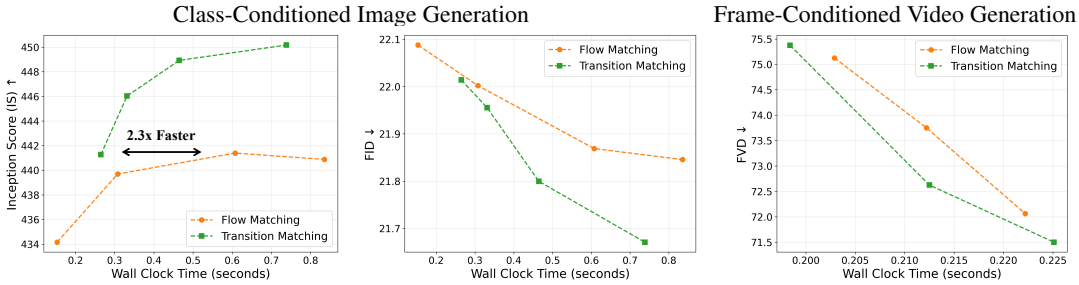

Figure 4: **Quantitative Evaluation v.s. Wall Clock Time.** Comparison of FM (orange) and TM (green) in class-conditioned image generation (left, middle) and frame-conditioned video generation (right), plotted against wall clock time measured in seconds.

Effect of Mean Separation on Sampling. To better understand multimodal targets, we study how the minimal mean separation D_{min} in Cor. 2 governs the sampling dynamics. We set the target distribution p_1 to a Gaussian mixture following (Cardoso et al., 2024) with predefined means and identity covariances. To analyze the sampling dynamics, we investigate the difference latent distribution, $p(V|X_{t_n})$. At each timestep t_n, we sample the difference latent M times to get $\{\widetilde{V}_{t_n}^{(m)}\}_{m=1}^{M} \sim p_\theta(V \mid X_{t_n})$ using the TM flow head in (10). We visualize the empirical distribution by computing the histogram of the cosine similarity between each $\widetilde{V}_{t_n}^{(m)}$ and the conditional expectation $\mathbb{E}[V \mid X_{t_n}]$ obtained from the samples, and plotting it across timesteps along the trajectory.

We present the result in Fig. 3 for two settings with $D_{\text{min}} \in \{8, 45\}$. When $D_{\text{min}} = 8$ (top row), the distribution exhibits multimodal behavior until $t_n = 0.25$, showing weak indication of the unimodal distribution approximation. For larger $D_{\text{min}} = 45$ (bottom row), Cor. 2 yields a tighter bound, and thus the upper bound approximation error at a fixed timestep t_n is smaller than for lower D_{min}. Empirically, we observe the cosine similarities clustering near 1 with non-negligible variance at earlier t_n. We present additional results analyzing the effect of target variance in §E.

KL Divergence Comparison. Building on the analysis, we now compare the KL divergence of FM and TM. First, Lem. 1 shows that once a trajectory enters the good region of a mode with sufficient margin, the dynamics can be locally approximated by the unimodal case in §3. Specifically, if we define $\mathcal{G}_t(r, \rho^*) \triangleq \left\{x : \|x - t\mu_{k_t(x)}\| \leq r \text{ and } \rho_t(x) \geq \rho^*\right\}$ for some r, ρ^*, then $\mathbb{P}(X_t \notin \mathcal{G}_t(r, \rho^*))$ decreases to zero exponentially as D_{min} increases. Then for $x \in \mathcal{G}(r, \rho^*)$, the Gaussian mixture can be closely approximated by a single unimodal component, with relative error that decays exponentially as r decreases or ρ^* increases. In this regime, the sampling dynamics approximate the unimodal Gaussian case described in §3, where Thm. 1 shows that the stochastic updates of TM better preserve the target variance and outperform FM. Formally, we show that the advantage of TM in the unimodal case extends to the mixture of Gaussians setting.

Theorem 2. *For any $M \in \{0, \ldots, N-1\}$ and $\beta \in \left(0, \frac{1}{2}\right)$, set $r = \beta t_M D_{\text{min}}$ and $\rho^* = (1 - 2\beta) t_M D_{\text{min}}$. Suppose $\widetilde{X}_{t_M}^{(\cdot)} \in \mathcal{G}_{t_M}(r, \rho^*)$, where $(\cdot)$ is either FM or TM, and r and ρ^*. Then,*

$$\text{KL}(p_1^{\text{TM}} \| p_1) < \text{KL}(p_1^{\text{FM}} \| p_1) - \gamma,$$

where γ can be arbitrarily close to 0, and $p_1^{(\cdot)}$ are the marginal distributions of the final iterates.

The proof of the theorem presented in §A.7 shows that when mixture components are sufficiently well separated (i.e., when D_{min} is large), trajectories that enter the good region of a mode remain confined within it with high probability throughout the remaining iterations. As a consequence, the subsequent sampling dynamics effectively follow the unimodal Gaussian case analyzed in §3, where TM provably achieves a smaller KL divergence than FM.

5 Empirical Results

We validate our theoretical insights on a synthetic mixture of Gaussians (§5.1), and evaluate them on *class-conditioned image generation* (§5.2) and *frame-conditioned video generation* (§5.3). In Appendix, we present implementation details in §D and additional results: diffusion-based model comparisons and qualitative results for image and video generation—in §E.

5.1 GAUSSIAN SYNTHETIC DATASET

To validate Thm. 2, we consider the bimodal case ($K = 2$) and evaluate performance under different target mean separations D_{min} in Cor. 2. We place the two component means on a circle of fixed radius at polar angles $\theta = \pm\frac{2\pi}{9}$ and $\theta = \pm\frac{4\pi}{9}$ with isotropic covariances $\sigma_k = 0.1$. This results to the case where $D_{\min} = 2\sin\left(\frac{2\pi}{9}\right)$ and $D_{\min} = 2\sin\left(\frac{4\pi}{9}\right)$, respectively.

Results. Fig. 5 presents the Monte Carlo estimates of the KL divergence between samples generated by FM and TM and the target mixture, plotted against wall clock time. Consistent with Thm. 2, TM attains lower KL divergence than FM in the low-step regime (wall clock time < 0.002) for both separations $D_{\min}$ ($\theta = \pm 2\pi/9$, green; $\theta = \pm 4\pi/9$, orange). Moreover, under fixed compute (red dotted line), the performance gap between FM and TM is greater for the larger $D_{\min}$, indicating that stronger local unimodality associated with larger $D_{\min}$ yields a structural advantage for TM via better variance preservation, and FM requires more steps to match TM.

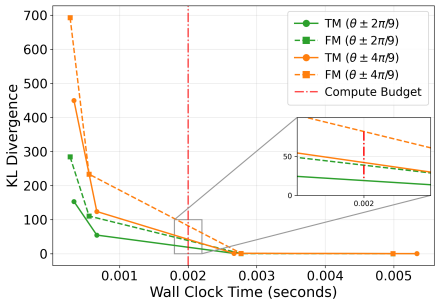

Figure 5: **KL Divergence for Mixture of Gaussians Target.** TM shows lower KL divergence than FM as the modes are more separated (red curve, larger D_{min}).

5.2 CLASS-CONDITIONED IMAGE GENERATION

We evaluate FM and TM on $50,000$ images from the test set and report Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017) to assess sample quality and diversity, where FID is computed between generated samples and the ground truth test set. For FM, we vary the number of sampling steps $N \in \{16, 32, 64, 96\}$, while for TM, we fix $N = 16$, and scale $S \in \{2, 4, 8, 16\}$. Quantitative evaluations are reported for each setting against wall clock time for generating a 256×256 image.

Results. Fig. 4 (left, middle) presents quantitative results under varying compute budgets. Consistent with the trend observed in §5.1, TM outperforms FM under a fixed compute budget across both evaluation metrics, achieving a Pareto front. Notably, TM with $S = 2$ achieves comparable IS to FM with $N = 64$, yielding a $2.3\times$ wall clock time speedup (black arrow).

5.3 FRAME-CONDITIONED VIDEO GENERATION

For video generation, we report Fréchet Video Distance (FVD) (Unterthiner et al., 2018), computed between generated and corresponding ground-truth videos, to assess temporal consistency and perceptual realism. Following the image generation experiment, we fix $N = 16$ for TM and scale $S \in \{12, 16, 20\}$, while N for FM is chosen to ensure comparable compute between the two methods. We report quantitative results along with the wall clock time required to generate a single video.

Results. The quantitative results are presented in Fig. 4 (right). Consistent with the trends observed in the mixture of Gaussians and class-conditioned image generation experiments, TM outperforms FM under a matched compute budget, achieving a more favorable quality–compute Pareto front.

6 CONCLUSION

In this work, we provided a theoretical results and analysis that explain when and why Transition Matching (TM) outperforms Flow Matching (FM). In unimodal Gaussian targets, we proved that TM achieves strictly lower KL divergence than FM given a same number of steps and converges faster under fixed compute. However, the distinction between TM and FM vanishes as the target variance approaches zero. We extend the analysis to Gaussian mixtures, identifying locally unimodal regimes in which the sampling dynamics approximate the unimodal case, where TM outperforms FM. Together, these results establish conditions under which TM should be preferred: targets with well-separated modes and non-negligible variances. Empirical studies on controlled Gaussian settings confirm the theoretical predictions, while experiments on real-world image and video generation further validated the practical impact.

ACKNOWLEDGEMENTS

We thank Kyeongmin Yeo for providing constructive feedback on the analysis of the unimodal Gaussian setup. We are also grateful to Tao Yu and Kaan Özkara for maintaining the server clusters and offering valuable advice on neural network optimization behaviors. This work was supported by the National Research Foundation of Korea (NRF) (RS-2026-25486000); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (RS-2024-00399817, RS-2025-25441313, RS-2025-25443318), funded by the Korean government (MSIT); the Industrial Technology Innovation Program (RS-2025-02317326), funded by the Korean government (MOTIE); the National Supercomputing Center (KSC-2025-CRE-0475); and the DRB-KAIST SketchTheFuture Research Center.

REFERENCES

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

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *NeurIPS*, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 2024.

Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. In *ICLR*, 2024.

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, 2024. In *Int. Conf. Machine. Learning.*, 2024.

Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. *URL https://diffusionflow. github. io*, 2024.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *NeurIPS*, 2024.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *ICLR*, 2025.

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.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky TQ Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. In *ICLR*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NIPS*, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching, 2024. In *Int. Conf. Machine. Learning.*, 2025.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal diagonal covariance matching. *arXiv preprint arXiv:2406.10808*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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

Gleb Ryzhakov, Svetlana Pavlova, Egor Sevriugov, and Ivan Oseledets. Explicit flow matching: On the theory of flow matching algorithms with applications. *arXiv preprint arXiv:2402.03232*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016.

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *arXiv preprint arXiv:2310.12395*, 2023.

Neta Shaul, Ricky TQ Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *Int. Conf. Machine. Learning.*, 2023.

Neta Shaul, Uriel Singer, Itai Gat, and Yaron Lipman. Transition matching: Scalable and flexible generative modeling. *arXiv preprint arXiv:2506.23589*, 2025.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021a.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Int. Conf. Machine. Learning.*, 2025.

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*, 2021b.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025.

Yilun Xu, Shangyuan Tong, and Tommi Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

APPENDIX

A PROOFS OF MAIN RESULTS

A.1 PROOF OF THM. 1: GAUSSIAN TARGET DISTRIBUTION

Theorem 3. *Let $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \mathcal{N}(\mu, \sigma^2 I_d)$ with $\sigma > 0$ be independent Gaussian vectors in $\mathbb{R}^d$. Consider the discretization of $[0,1]$ as $t_n = n\Delta t$, $\Delta t = 1/N$ with $N > 1$, and let $S > 1$ be the number of inner ODE steps for TM in (10). If FM and TM iterates are updated according to (6) and (10), respectively, then*

$$\mathrm{KL}\left(p_1^{\mathrm{TM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) < \mathrm{KL}\left(p_1^{\mathrm{FM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right),$$

where p_1^{TM} and p_1^{FM} are the marginal distributions of $\widetilde{X}_{t_N=1}^{\mathrm{TM}}$ and $\widetilde{X}_{t_N=1}^{\mathrm{FM}}$, respectively.

Proof. For any $t \in [0,1]$, let the corresponding point on the linear trajectory from X_0 to X_1 be $X_t = (1-t)X_0 + tX_1$, and let $V = X_1 - X_0$ be the difference latent . Since $X_0 \perp X_1$, and X_t is a convex combination of two independent Gaussian random variables, from the linearity of covariance, we have,

$$\mathrm{Cov}(X_t) \triangleq \mathbb{E}\left[(X_t - \mathbb{E}[X_t])(X_t - \mathbb{E}[X_t])^\top\right] = B(t) I_d, \quad \text{where } B(t) \triangleq (1-t)^2 + \sigma^2 t^2, \tag{14}$$

$$\text{and,} \quad \mathrm{Cov}(X_t, V) \triangleq \mathbb{E}\left[(X_t - \mathbb{E}[X_t])(V - \mathbb{E}[V])^\top\right] = A(t) I_d, \quad \text{where } A(t) \triangleq t(1+\sigma^2) - 1. \tag{15}$$

Using some algebra, it can be shown that

$$A(t)^2 + \sigma^2 = (1+\sigma^2) B(t). \tag{16}$$

Since $X_0 \perp X_1$, the stacked random variable $(X_0, X_1) \in \mathbb{R}^{2d}$ is jointly Gaussian. Consequently, (X_t, V) is also jointly Gaussian, since it is a linear transformation of (X_0, X_1). Using Lemma 4 and (16) above, $V | X_t = x$ is also Gaussian-distributed, and we have,

$$\mathbb{E}[V | X_t = x] = \mu + k(t)(x - \mu t), \quad \mathrm{Cov}(V | X_t = x) = \tau^2(t) I_d, \tag{17}$$
$$\text{where } k(t) \triangleq \frac{A(t)}{B(t)} \quad \text{and} \quad \tau^2(t) \triangleq \frac{\sigma^2}{B(t)}.$$

KL Divergence for FM steps. From the expression in (6) and the fact that perfect minimization of (4) will ensure $v_{t_n}^\theta = \mathbb{E}[V | \widetilde{X}_{t_n}^{\mathrm{FM}}]$ (5), the deterministic update of the FM Euler step can be written recursively as

$$\widetilde{X}_{t_{n+1}}^{\mathrm{FM}} = a_n \widetilde{X}_{t_n}^{\mathrm{FM}} + b_n, \quad \text{where, } a_n \triangleq 1 + \Delta t k(t_n) \text{ and } b_n \triangleq \Delta t\big(\mu - k(t_n)\mu t_n\big) \tag{18}$$

Let $m_n = \mathbb{E}[\widetilde{X}_{t_n}^{\mathrm{FM}}]$. Then,

$$m_{n+1} = a_n m_n + b_n, \qquad m_0 = 0.$$

By induction, $m_n = \mu t_n$ for all n since

$$m_{n+1} = (1 + \Delta t k(t_n))\mu t_n + \Delta t(\mu - k(t_n)\mu t_n) = \mu t_n + \Delta t \mu = \mu t_{n+1}. \tag{19}$$

Therefore, the mean follows the interpolation path and leads to $m_N = \mu$.

Furthermore, from (18), the covariance evolves as

$$\mathrm{Cov}\left(\widetilde{X}_{t_{n+1}}^{\mathrm{FM}}\right) = a_n^2 \,\mathrm{Cov}\left(\widetilde{X}_{t_n}^{\mathrm{FM}}\right). \tag{20}$$

In other words, let $\mathrm{Cov}\left(\widetilde{X}_{t_n}^{\mathrm{FM}}\right) = s_n^{\mathrm{FM}} I_d$, and the evolution of s_N^{FM} is given by,

$$s_{n+1}^{\mathrm{FM}} = a_n^2 s_n^{\mathrm{FM}}, \qquad s_0^{\mathrm{FM}} = 1. \quad \Rightarrow \quad s_N^{\mathrm{FM}} = \prod_{n=0}^{N-1} \big(1 + \Delta t k(t_n)\big)^2. \tag{21}$$

On the other hand, the true variance, i.e., $B(t_n)$ defined in (14) can be written as a recursion as,

$$\begin{aligned} B(t_{n+1}) &= B(t_n) + 2\Delta t A(t_n) + (\Delta t)^2(1+\sigma^2) \\ &\overset{\text{(i)}}{=} B(t_n) + 2\Delta t k(t_n) B(t_n) + (\Delta t)^2 \left(k(t_n)^2 B(t_n) + \frac{\sigma^2}{B(t_n)} \right) \\ &= (1+\Delta t k(t_n))^2 B(t_n) + (\Delta t)^2 \frac{\sigma^2}{B(t_n)}, \end{aligned} \tag{22}$$

where (i) makes use of the notation of $k(t_n)$ along with (16). Define the ratio of the variance of the sample obtained from FM Euler step to the true variance of the interpolant as $r_n^{\mathrm{FM}} \triangleq s_n^{\mathrm{FM}}/B(t_n)$. From (21) and (22),

$$r_{n+1}^{\mathrm{FM}} = w_n r_n^{\mathrm{FM}}, \qquad r_0^{\mathrm{FM}} = 1, \tag{23}$$

where

$$w_n \triangleq \frac{(1+\Delta t k(t_n))^2}{(1+\Delta t k(t_n))^2 + (\Delta t)^2 \sigma^2 / B(t_n)^2} \in (0,1). \tag{24}$$

Hence, the *variance contraction factor*, r_n^{FM} strictly decreases and is always less than 1, which means,

$$s_N^{\mathrm{FM}} < B(1) = \sigma^2. \tag{25}$$

With matching means (19) and isotropic covariances, the KL divergence (5) between the generated samples at $t_N = 1$ and the target is given by

$$\mathrm{KL}\left(p_{t_N}^{\mathrm{FM}} \| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \frac{d}{2}\left(\frac{s_N^{\mathrm{FM}}}{\sigma^2} - 1 - \log \frac{s_N^{\mathrm{FM}}}{\sigma^2} \right) > 0. \tag{26}$$

In Cor. 1 (proved in §A.2), we show that this KL divergence converges to 0 as $N \to \infty$ in $\mathcal{O}\left(\frac{1}{N^2}\right)$.

KL divergence for TM steps. As seen from (10), given $X_0 \sim \mathcal{N}(0, I_d)$, for every outer iteration $n \in [N-1]$, TM draws a sample from the conditional distribution of the *difference latent*, $p(V|\widetilde{X}_{t_n}^{\mathrm{TM}})$, by iterating over S inner Euler steps. Analogous to (18), the TM updates can also be written as

$$\widetilde{X}_{t_{n+1}}^{\mathrm{TM}} = a_n \widetilde{X}_{t_n}^{\mathrm{TM}} + b_n + \eta_n, \qquad \eta_n = \Delta t \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}\big(0, c_S \tau^2(t_n) I_d\big), \tag{27}$$

where the additional variance term η_n arises from the stochasticity of the updates as a result of sampling from $p(V|\widetilde{X}_{t_n}^{\mathrm{TM}})$, instead of using $\mathbb{E}[V|\widetilde{X}_{t_n}^{\mathrm{TM}}]$ as the update direction. If the sampling of $V|\widetilde{X}_{t_n}^{\mathrm{TM}}$ had been perfect, then $\mathrm{Cov}(\eta_n) = \tau^2(t_n) I_d$ from (17). For instance, this is the case when the number of inner iterations $S \to \infty$. However, when a *finite* number of inner Euler steps is used to simulate the inner ODE, and consequently, approximately sample $V|\widetilde{X}_{t_n}^{\mathrm{TM}}$, it introduces an additional scaling factor akin to (23) – the effect of which is captured using c_S. This mirrors the FM case, as the target distribution, i.e., (17), is also unimodal Gaussian. In other words,

$$\mathrm{Cov}\left(\widetilde{V}_{t_n} \mid \widetilde{X}_{t_n}^{\mathrm{TM}}\right) = c_S \tau^2(t_n) I_d, \qquad 0 < c_S \le 1, \tag{28}$$

where $\widetilde{V}_{t_n}$ is defined as in (10). Note that $c_S \to 1$ as $S \to \infty$ as shown in Cor. A.2, and the inner ODE recovers the true $p(V|\widetilde{X}_{t_n}^{\mathrm{TM}})$.

Similar to the FM case in Equation. 19, the mean follows the interpolation path. That is, overloading the notation m_n, and denoting $m_n = \mathbb{E}[\tilde{X}_{t_n}^{\mathrm{TM}}]$, we see from induction,

$$m_{n+1} = (1+\Delta t k(t_n))\mu t_n + \Delta t(\mu - k(t_n)\mu t_n) = \mu t_{n+1}.$$

Let us denote $\mathrm{Cov}(\widetilde{X}_{t_n}^{\mathrm{TM}}) = s_n^{\mathrm{TM}} I_d$. From (27),

$$s_{n+1}^{\mathrm{TM}} = a_n^2 s_n^{\mathrm{TM}} + (\Delta t)^2 c_S \tau^2(t_n) = a_n^2 s_n^{\mathrm{TM}} + (\Delta t)^2 c_S \frac{\sigma^2}{B(t_n)}. \tag{29}$$

With $r_n^{\mathrm{TM}} \triangleq s_n^{\mathrm{TM}}/B(t_n)$ and using (22) and (29),

$$r_{n+1}^{\mathrm{TM}} = \frac{(1+\Delta t k(t_n))^2 r_n^{\mathrm{TM}} + (\Delta t)^2 c_S \sigma^2 / B(t_n)^2}{(1+\Delta t k(t_n))^2 + (\Delta t)^2 \sigma^2 / B(t_n)^2} = w_n r_n^{\mathrm{TM}} + (1-w_n) c_S, \qquad r_0^{\mathrm{TM}} = 1, \tag{30}$$

where w_n is as defined in (24). Combining (23) and (30), we get,

$$r_n^{\mathrm{FM}} < r_n^{\mathrm{TM}} \leq 1 \quad \Longleftrightarrow \quad s_n^{\mathrm{FM}} < s_n^{\mathrm{TM}} \leq B(t_n) \quad \text{for all } n. \tag{31}$$

Since the means are the same for $t_N = 1$, so the KL divergence between the target distribution and the distribution of $\widetilde{X}_{t_N}^{\mathrm{TM}}$ is given by

$$\mathrm{KL}\left(p_{t_N}^{\mathrm{TM}} \| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \frac{d}{2}\left(\frac{s_N^{\mathrm{TM}}}{\sigma^2} - 1 - \log\frac{s_N^{\mathrm{TM}}}{\sigma^2}\right). \tag{32}$$

Since $\phi(x) = \frac{d}{2}(x - 1 - \log x)$ is strictly decreasing on $(0, 1]$, (26) and (32) together with the fact $t_N = 1$, imply,

$$\mathrm{KL}\left(p_1^{\mathrm{TM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) < \mathrm{KL}\left(p_1^{\mathrm{FM}} \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right). \tag{33}$$

This concludes the proof. □

A.2 Proof of Corollary 1: KL divergence decay rate for FM and TM

Corollary. [KL divergence decay rate for FM and TM] *Under the setting of Thm. 1 with the discretization of* $[0, 1]$ *given by* $t_n = n\Delta t$, $\Delta t = 1/N$, *the covariance of the FM iterates, given by* $s_N^{\mathrm{FM}} = \prod_{n=0}^{N-1}(1 + \Delta t k(t_n))^2$, *satisfies* $s_N^{\mathrm{FM}} = \sigma^2 + \mathcal{O}\left(N^{-1}\right)$ *in the asymptotic limit of* $N \to \infty$. *Consequently, we have,*

$$\mathrm{KL}\left(p_1^{\mathrm{FM}}(N) \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \mathcal{O}\left(\frac{1}{N^2}\right).$$

Moreover, the TM iterates obtained according to (10) satisfy

$$\mathrm{KL}\left(p_1^{\mathrm{TM}}(N, S) \middle\| \mathcal{N}(\mu, \sigma^2 I_d)\right) = \mathcal{O}\left(\frac{1}{N^2 S^2}\right) \to 0,$$

Proof. **Asymptotic decay rate for FM**. We begin by taking the logarithm of s_N^{FM}, i.e.,

$$\log s_N^{\mathrm{FM}} = 2\sum_{n=0}^{N-1} \log\left(1 + \Delta t k(t_n)\right).$$

Using the first-order Taylor expansion, $\log(1 + x) = x + \mathcal{O}(x^2)$, and retaining only the leading term,

$$\log s_N^{\mathrm{FM}} = 2\sum_{n=0}^{N-1}\left[\Delta t k(t_n) + \mathcal{O}\left(\frac{k(t_n)^2}{N^2}\right)\right] = 2\sum_{n=0}^{N-1}\Delta t k(t_n) + \mathcal{O}\left(\frac{1}{N}\right).$$

The summation can be viewed as a Riemann sum, and subsequently approximated as an integral:

$$\log s_N^{\mathrm{FM}} = 2\int_0^1 k(t)dt + \mathcal{O}\left(\frac{1}{N}\right).$$

Since $k(t)$ is continuously differentiable in $[0, 1]$, the approximation error of this integral is also $\mathcal{O}(N^{-1})$.

Using (16), we have $k(t) = A(t)/B(t)$ and $B'(t) = 2A(t)$, which yield,

$$2\int_0^1 k(t)dt = \int_0^1 \frac{B'(t)}{B(t)}dt = \log B(1) - \log B(0).$$

Since $B(0) = 1$ and $B(1) = \sigma^2$, it follows that

$$\log s_N^{\mathrm{FM}} = \log\sigma^2 + \mathcal{O}\left(\frac{1}{N}\right).$$

Exponentiating both sides gives

$$s_N^{\mathrm{FM}} = \exp\left(\log\sigma^2 + \mathcal{O}\left(\frac{1}{N}\right)\right) = \sigma^2\left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right) = \sigma^2 + \mathcal{O}\left(\frac{1}{N}\right). \tag{34}$$

Now consider the KL divergence between $p_{t_N}^{\mathrm{FM}} = \mathcal{N}(\mu, s_N^{\mathrm{FM}} I_d)$ and the target $\mathcal{N}(\mu, \sigma^2 I_d)$, given by:

$$\mathrm{KL} = \frac{d}{2}\left(\frac{s_N^{\mathrm{FM}}}{\sigma^2} - 1 - \log\left(\frac{s_N^{\mathrm{FM}}}{\sigma^2}\right)\right). \tag{35}$$

Let us denote $\varepsilon_N \triangleq (s_N^{\mathrm{FM}} - \sigma^2)/\sigma^2$. Then from (34), $\epsilon_N = \mathcal{O}(N^{-1})$. Using the Taylor series expansion of $\log(1+x)$, we get,

$$\mathrm{KL} = \frac{d}{2}\left(\varepsilon_N - \log(1+\varepsilon_N)\right) = \frac{d}{2}\left(\varepsilon_N - (\varepsilon_N - \tfrac{1}{2}\varepsilon_N^2 + \mathcal{O}(\varepsilon_N^3))\right) = \frac{d}{4}\varepsilon_N^2 + \mathcal{O}(\varepsilon_N^3) = \mathcal{O}\left(\frac{1}{N^2}\right).$$

This finishes the derivation of the convergence rate for FM.

Asymptotic decay rate for TM. The derivation for the convergence rate of TM relies on the fact that the inner ODE iterations for sampling the difference latent are FM iterations themselves. Hence, we require an expression for how the variance contraction factor, r_N^{FM} from (23) depends on the number of iterations, N. This was unnecessary for proving Thm 1, but is needed to bound the error accumulation from the inner TM iterations.

Firstly, recall from (28) that the true covariance of the difference latent is $\mathrm{Cov}\left(\widetilde{V}_{t_n} \mid \widetilde{X}_{t_n}^{\mathrm{TM}}\right) = c_S \tau^2(t_n) I_d$, where $\tau_n^2(t) = \frac{\sigma^2}{B(t_n)}$, and the variance contraction factor $c_S \leq 1$ plays the same role as r_n^{TM} from (23) in the FM analysis. Hence, for the remainder of this proof, we bound r_N^{FM}, while an analogous bound holds for c_S, with N replaced by S. Note that c_S does not depend n, since once $\widetilde{X}_{t_n}^{\mathrm{TM}}$ is fixed, c_S characterizes the variance contraction of the inner ODE iterations, and hence, depends only on S, the total number of inner ODE steps.

The evolution of r_n^{FM} is specified by (23) and (24). Therefore, $r_N^{\mathrm{FM}} = \prod_{n=0}^{N-1} w_n$. Recall the notation $a_n = (1 + \Delta t k(t_n))$, and denote $g_n \triangleq \sigma^2/B(t_n)^2$. Then,

$$\begin{aligned} w_n = \frac{a_n^2}{a_n^2 + (\Delta t)^2 g_n} \implies \log w_n &= \log a_n^2 - \log\left(a_n^2 + (\Delta t)^2 g_n\right) \\ &= -\log\left(1 + \frac{(\Delta t)^2 g_n}{a_n^2}\right). \end{aligned} \tag{36}$$

To expand a_n^{-2}, consider the Taylor series expansion, $(1+x)^{-2} = 1 - 2x + 3x^2 + \mathcal{O}(x^3)$ for $|x| < 1$. Here, $x = \Delta t k(t_n)$ and as $k(t_n) = A(t_n)/B(t_n)$ is bounded, $|\Delta t k(t_n)| = |k(t_n)/N| < 1$ for large enough N. Therefore,

$$\begin{aligned} a_n^{-2} &= (1 + \Delta t k(t_n))^{-2} = 1 - 2\Delta t k(t_n) + 3(\Delta t)^2 k(t_n)^2 + \mathcal{O}(\Delta t^3) \\ \frac{(\Delta t)^2 g_n}{a_n^2} &= (\Delta t)^2 g_n - 2(\Delta t)^3 g_n k(t_n) + \mathcal{O}(\Delta t^4). \end{aligned} \tag{37}$$

Next, since $\log(1+x) = x - \frac{x^2}{2} + \mathcal{O}(x^3)$ for $|x| < 1$, and $(\Delta t)^2 g_n / a_n^2 < 1$ for large enough N, using (36), we get,

$$\begin{aligned} \log w_n = -\log\left(1 + \frac{(\Delta t)^2 g_n}{a_n^2}\right) &= -\frac{(\Delta t)^2 g_n}{a_n^2} + \mathcal{O}(\Delta t^4) \\ &= -(\Delta t)^2 g_n + 2(\Delta t)^3 g_n k(t_n) + \mathcal{O}(\Delta t^4) \\ &= -\frac{\sigma^2 (\Delta t)^2}{B(t_n)^2} + \mathcal{O}(\Delta t^3). \end{aligned} \tag{38}$$

Using this with (23) yields,

$$\log r_N^{\mathrm{FM}} = \sum_{n=0}^{N-1} \log w_n = -\sigma^2 \sum_{n=0}^{N-1} \frac{(\Delta t)^2}{B(t_n)^2} + \mathcal{O}\left(\frac{1}{N^2}\right) \overset{\text{(i)}}{=} -\frac{\sigma^2}{N} \int_0^1 \frac{dt}{B(t)^2} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{39}$$

where in (i), we use $\Delta t = 1/N$ and approximate the summation with a finite integral for large N. Since $B(t)$ is bounded in $[0, 1]$, the finite integral is also bounded, and is a function of σ. Denote

$P(\sigma) \triangleq \sigma^2 \int_0^1 B(t)^{-2} dt$. While $P(\sigma)$ can be evaluated exactly, it is not required for the purposes of this derivation, and can simply be treated as a constant. Taking exponentials in (39), and using $e^x = 1 + x + \mathcal{O}(x^2)$, this gives,

$$r_N^{\mathrm{FM}} = \exp\left(-\left(\frac{P(\sigma)}{N}\right) + \mathcal{O}\left(\frac{1}{N^2}\right)\right) = 1 - \frac{P(\sigma)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{40}$$

Now that we are able to show the asymptotic dependence of r_N^{FM} on N, a similar result holds true for the inner ODE iterations in TM, and is given by

$$c_S = 1 - \frac{P(\tau(t_n))}{S} + \mathcal{O}\left(\frac{1}{S^2}\right), \tag{41}$$

where given $\widetilde{X}_{t_n}^{\mathrm{TM}}$, $\tau(t_n)$ is as specified in (17).

From (30), the evolution of the variance contraction ratio for TM, r_n^{TM}, is given by

$$r_{n+1}^{\mathrm{TM}} = w_n r_n^{\mathrm{TM}} + (1 - w_n) c_S, \quad r_0^{\mathrm{TM}} = 1, \text{ and, } w_n = \frac{a_n^2}{a_n^2 + (\Delta t)^2 \sigma^2 / B(t_n)^2}. \tag{42}$$

Since r_{n+1}^{TM} is a convex combination of the previous ratio, r_n^{TM}, and c_S, which is independent of n, the evolution for $N \geq 1$ can be expressed as,

$$r_n^{\mathrm{TM}} = c_S + (1 - c_S) \prod_{n=0}^{N-1} w_n. \tag{43}$$

From (41), we can express $c_S = 1 + \delta_S$, where $\delta_S = \mathcal{O}\left(S^{-1}\right)$. Therefore, using this,

$$s_{N,S}^{\mathrm{TM}} = r_N^{\mathrm{TM}} B(1) = \sigma^2 \left(c_S + (1 - c_S) r_N^{\mathrm{FM}}\right) = \sigma^2 \left(1 + \delta_S (1 - r_N^{\mathrm{FM}})\right). \tag{44}$$

Hence, proceeding as done after (35), the relative variance error is

$$\varepsilon_{N,S} \triangleq \frac{s_{N,S} - \sigma^2}{\sigma^2} = \delta_S \left(1 - r_N^{\mathrm{FM}}\right), \tag{45}$$

and the KL divergence is then give by,

$$\begin{aligned} \mathrm{KL} &= \frac{d}{2}\left(\varepsilon_{N,S} - \log\left(1 + \varepsilon_{N,S}\right)\right) = \frac{d}{4}\varepsilon_{N,S}^2 + \mathcal{O}\left(\varepsilon_{N,S}^3\right) = \mathcal{O}\left(\delta_S^2 \left(1 - r_N^{\mathrm{FM}}\right)^2\right) \\ &= \mathcal{O}\left(\frac{\left(1 - r_N^{\mathrm{FM}}\right)^2}{S^2}\right) \overset{\text{(i)}}{=} \mathcal{O}\left(\frac{1}{N^2 S^2}\right), \end{aligned} \tag{46}$$

where (i) utilizes the asymptotic expression for r_n^{FM} from (40) to get,

$$\left(1 - r_N^{\mathrm{FM}}\right)^2 = \left(\frac{P(\sigma)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)\right)^2 = \mathcal{O}\left(\frac{1}{N^2}\right). \tag{47}$$

This completes the proof.

□

A.3 Proof of Proposition 1: Unimodal approximation of difference latent distribution

Proposition. [Unimodal approximation of difference latent distribution] *Let $X_0 \sim \mathcal{N}(0, I_d)$ be sampled from the standard Gaussian distribution, and let*

$$X_1 \sim \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j^2 I_d)$$

be sampled from a Gaussian mixture with K components, where $\pi_j > 0, \sigma_j > 0$, and $\sum_{j=1}^{K} \pi_j = 1$. Assume that X_0 and X_1 are independent. For any fixed $t \in (0, 1]$, define the linear interpolation

$X_t = (1-t)X_0 + tX_1$, and let $B_t(j) = (1-t)^2 + t^2\sigma_j^2$ be the variance conditioned on X_1 being drawn from j^{th} component (i.e., $Z = j$). Given $x \in \mathbb{R}^d$, let $k_t(x) \in \operatorname{argmin}_j \|x - t\mu_j\|$ denote the index of the nearest path mean (with ties are broken arbitrarily), and let $D_t(x) \triangleq \|x - t\mu_{k_t(x)}\|$ be the distance to it. Furthermore, define the associated margin,

$$\rho_t(x) \triangleq \min_{j \neq k_t(x)} \left(\|x - t\mu_j\| - \|x - t\mu_{k_t(x)}\| \right),$$

to denote the gap between the distances to the closest and the second-closest path means. Then, for $V = X_1 - X_0$,

$$\|p(V|X_t = x) - p(V|X_t = x, Z = k_t(x))\|_{\text{TV}} \leq \left(\frac{1}{\pi_{k_t(x)}} - 1\right)\left(\frac{B_t^{\max}}{B_t^{\min}}\right)^{\frac{d}{2}} \exp\left(\frac{D_t^2(x)}{2}\left(\frac{1}{B_t^{\min}} - \frac{1}{B_t^{\max}}\right) - \frac{\rho_t^2(x)}{2B_t^{\max}}\right), \tag{48}$$

where $B_t^{\min} \triangleq \min_j B_t(j)$ and, $B_t^{\max} \triangleq \max_j B_t(j)$

Proof. Conditioned on being sampled from the jth Gaussian component, we have,

$$X_t | Z = j \sim \mathcal{N}\left(t\mu_j, B_t(j) I_d\right). \tag{49}$$

For a Gaussian mixture model, *posterior responsibility* refers to the posterior probability that a given data point was generated from a specific Gaussian component. Denote the posterior responsibilities as

$$w_t(x, j) \triangleq \mathbb{P}(Z = j | X_t = x) = \frac{\pi_j B_t(j)^{-d/2} \exp\left(-\|x - t\mu_j\|^2 / 2B_t(j)\right)}{\sum_{\ell=1}^K \pi_\ell B_t(\ell)^{-d/2} \exp\left(-\|x - t\mu_\ell\|^2 / 2B_t(\ell)\right)} \tag{50}$$

Marginalizing over the probability of the components, we have,

$$p(V|X_t = x) = \sum_j w_t(x, j) p(V|X_t = x, Z = j)$$

Utilizing $\sum_\ell w_t(x, \ell) = 1$ and triangle inequality,

$$\begin{aligned}
&\|p(V|X_t = x) - p(V|X_t = x, Z = k)\|_{\text{TV}} \\
&\quad = \left\| \sum_j w_t(x, j) p(V|X_t = x, Z = j) - p(V|X_t = x, Z = k) \right\|_{\text{TV}} \\
&\quad \leq \sum_{j \neq k} w_t(x, j) \, \|p(V|X_t = x, Z = j) - p(V|X_t = x, Z = k)\|_{\text{TV}} \\
&\quad \overset{\text{(i)}}{\leq} \sum_{j \neq k} w_t(x, j) = 1 - w_t(x, k).
\end{aligned} \tag{51}$$

Here, (i) follows from the fact that TV distance is always upper bounded by 1. Therefore, it suffices to upper bound (51). To this end, for $k \equiv k_t(x)$, define the *responsibility ratio* as,

$$r_{j|k}(x) \triangleq \frac{w_t(x, j)}{w_t(x, k)} = \frac{\pi_j}{\pi_k} \left(\frac{B_t(x, k)}{B_t(x, j)}\right)^{\frac{d}{2}} \exp\left(-\frac{\|x - t\mu_j\|^2}{2B_t(x, j)} + \frac{\|x - t\mu_k\|^2}{2B_t(x, k)}\right). \tag{52}$$

We also have,

$$w_t(x, k) = \frac{1}{1 + \sum_{j \neq k} r_{j|k}(x)} \implies 1 - w_t(x, k) = \frac{\sum_{j \neq k} r_{j|k}(x)}{1 + \sum_{j \neq k} r_{j|k}(x)} \overset{\text{(i)}}{\leq} \sum_{j \neq k} r_{j|k}(x), \tag{53}$$

where (i) follows from the fact that the denominator $1 + \sum_{j \neq k} r_{j|k}(x)$ always exceeds 1. To upper bound $\sum_{j \neq k} r_{j|k}(x)$, note that the expression inside the $\exp(\cdot)$ in (52) is,

$$\begin{aligned}
-\frac{\|x - t\mu_j\|^2}{2B_t(x, j)} + \frac{D_t^2(x)}{2B_t(x, k)} &\leq -\frac{\|x - t\mu_j\|^2}{2B_t^{\max}} + \frac{D_t^2(x)}{2B_t^{\min}} \\
&\overset{\text{(i)}}{\leq} -\frac{\rho_t^2(x)}{2B_t^{\max}} + \frac{D_t^2(x)}{2}\left(\frac{1}{B_t^{\min}} - \frac{1}{B_t^{\max}}\right),
\end{aligned} \tag{54}$$

where to get (i), we add and subtract $\frac{D_t^2(x)}{2B_t^{\max}}$ and use the fact that $\|x - t\mu_j\| \geq D_t(x) \implies \|x - t\mu_j\|^2 - D_t^2(x) \geq \rho_t^2(x)$. Substituting this in (53) yields,

$$\sum_{j\neq k} r_{j|k}(x) \leq \exp\left(\frac{D_t^2(x)}{2}\left(\frac{1}{B_t^{\min}} - \frac{1}{B_t^{\max}}\right) - \frac{\rho_t^2(x)}{2B_t^{\max}}\right)\sum_{j\neq k}\frac{\pi_j}{\pi_k}\left(\frac{B_t(x,k)}{B_t(x,j)}\right)^{d/2}. \tag{55}$$

Upper bounding $\frac{B_t(x,k)}{B_t(x,j)} \leq \frac{B_t^{\max}}{B_t^{\min}}$ and noting that $\sum_{j\neq k}\frac{\pi_j}{\pi_k} = \frac{1}{\pi_{k_t(x)}} - 1$, completes the proof. □

A.4 Proof of Corollary 2: Effect of mode separation

Corollary. [Effect of mode separation] *Under the setting of Prop. 1, and* $\rho_t(x) \triangleq \min_{j\neq k_t(x)}\left(\|x - t\mu_j\| - \|x - t\mu_{k_t(x)}\|\right)$*, define minimal mean separation,* $D_{\min} \triangleq \min_{j\neq k}\|\mu_j - \mu_k\|$ *and the neighborhood* $\mathcal{E}_t \triangleq \left\{x \in \mathbb{R}^d : D_t(x) \leq \sqrt{B_t^{\min}}\right\}$*. Assuming* $B_t^{\max} = B_t^{\min}$*, for any* $x \in \mathcal{E}_t$*,*

$$\left\|p(V \mid X_t = x) - p(V \mid X_t = x, Z = k_t(x))\right\|_{\mathrm{TV}} \leq \left(\frac{1}{\pi_{k_t(x)}} - 1\right)\exp\left(2 - \frac{t^2 D_{\min}^2}{4B_t^{\max}}\right).$$

Proof. Starting from the original bound presented in Prop. 1,

$$\begin{aligned}&\|p(V|X_t = x) - p(V|X_t = x, Z = k_t(x))\|_{\mathrm{TV}}\\ &\qquad\leq \left(\frac{1}{\pi_{k_t(x)}} - 1\right)\left(\frac{B_t^{\max}}{B_t^{\min}}\right)^{\frac{d}{2}}\exp\left(\frac{D_t^2(x)}{2}\left(\frac{1}{B_t^{\min}} - \frac{1}{B_t^{\max}}\right) - \frac{\rho_t^2(x)}{2B_t^{\max}}\right).\end{aligned}$$

From the definition of the neighborhood, $D_t(x)$ is upper bounded by $\sqrt{B_t^{\min}}$, and the bound gives

$$\begin{aligned}&\left\|p(V \mid X_t = x) - p(V \mid X_t = x, Z = k_t(x))\right\|_{\mathrm{TV}} \leq\\ &\qquad\left(\frac{1}{\pi_{k_t(x)}} - 1\right)\left(\frac{B_t^{\max}}{B_t^{\min}}\right)^{\frac{d}{2}}\exp\left(\frac{1}{2}\left(1 - \frac{B_t^{\min}}{B_t^{\max}}\right) - \frac{\rho_t^2(x)}{2B_t^{\max}}\right).\end{aligned} \tag{56}$$

From the definition of $D_{\min}$ and $\rho_t(x)$, i.e.,

$$\rho_t(x) \triangleq \min_{j\neq k_t(x)}\left(\|x - t\mu_j\| - \|x - t\mu_{k_t(x)}\|\right), \quad D_{\min} \triangleq \min_{j\neq k}\|\mu_j - \mu_k\|,$$

an application of reverse triangle inequality yields for any $j \neq k_t(x)$,

$$\|x - t\mu_j\| \geq \|t\mu_j - t\mu_{k_t(x)}\| - \|x - t\mu_{k_t(x)}\| \geq tD_{\min} - D_t(x).$$

This implies, $\|x - t\mu_j\| - D_t(x) \geq tD_{\min} - 2D_t(x)$, and taking the minimum over $j \neq k_t(x)$ gives

$$\rho_t(x) \geq tD_{\min} - 2D_t(x) \geq tD_{\min} - 2\sqrt{B_t^{\min}},$$

where the second inequality follows as $D_t(x) \leq \sqrt{B_t^{\min}}$. Hence,

$$-\frac{\rho_t^2(x)}{2B_t^{\max}} \;\leq\; -\frac{\left(tD_{\min} - 2\sqrt{B_t^{\min}}\right)^2}{2B_t^{\max}}.$$

Additionally, from the inequality $(a - b)^2 \geq \frac{1}{2}a^2 - b^2$, the square term in the exponential gives $(tD_{\min} - 2\sqrt{B_t^{\min}})^2 \geq \frac{1}{2}t^2 D_{\min}^2 - 4B_t^{\min}$, which gives

$$-\frac{\left(tD_{\min} - 2\sqrt{B_t^{\min}}\right)^2}{2B_t^{\max}} \;\leq\; -\frac{t^2 D_{\min}^2}{4B_t^{\max}} + \frac{2B_t^{\min}}{B_t^{\max}}.$$

Plugging in (56), then for any $x \in \mathcal{E}_t$,

$$\begin{aligned}
&\left\|p(V \mid X_t = x) - p(V \mid X_t = x, Z = k_t(x))\right\|_{\mathrm{TV}} \\
\leq &\Big(\frac{1}{\pi_{k_t(x)}} - 1\Big)\Big(\frac{B_t^{\max}}{B_t^{\min}}\Big)^{d/2} \exp\left(\frac{1}{2}\Big(1 - \frac{B_t^{\min}}{B_t^{\max}}\Big) + \frac{2B_t^{\min}}{B_t^{\max}} - \frac{t^2 D_{\min}^2}{4B_t^{\max}}\right) \\
\approx &\Big(\frac{1}{\pi_{k_t(x)}} - 1\Big) \exp\left(2 - \frac{t^2 D_{\min}^2}{4B_t^{\max}}\right),
\end{aligned} \tag{57}$$

where the second approximation holds from the assumption that $B_t^{\max} = B_t^{\min}$. This completes the proof.

□

A.5 Local Unimodal Approximation for Gaussian Mixture

For Gaussian mixtures, after a short period, both FM or TM iterates fall into one component's basin. From that point onwards in time, the nearest mode dominates, and the future updates effectively follow the unimodal case. Consequently, we can invoke Thm. 1 for the remainder of the trajectory, and conclude that TM will outperform FM. The following Lemma bounds the probability with which a single mode dominates the local trajectory.

Lemma 1. *Fix $t \in (0, 1]$ and define the* **good region** *parameterized by (r, ρ^*) as,*

$$\mathcal{G}_t(r, \rho^*) \triangleq \left\{x : \|x - t\mu_{k_t(x)}\| \leq r \text{ and } \rho_t(x) \geq \rho^*\right\}, \tag{58}$$

where the nearest path mean, $k_t(x)$ and the margin, $\rho_t(x)$ are defined in Prop. 1. Then,

$$\mathbb{P}\left(X_t \notin \mathcal{G}_t(r, \rho^*)\right) \leq \exp\left(-\frac{r^2}{2B_t^{\min}}\right) + \exp\left(-\frac{(tD_{min} - \rho^*)^2}{8B_t^{\min}}\right), \tag{59}$$

where $D_{\min} \triangleq \min_{j \neq k} \|\mu_j - \mu_k\|$ (as defined in Cor. 2), and $B_t^{\min} \triangleq \min_j B_t(j)$.

Proof. From union bound and the law of total probability, we have,

$$\mathbb{P}\left(X_t \notin \mathcal{G}_t(r, \rho^*)\right) \leq \sum_j \pi_j \mathbb{P}\left(\|X_t - t\mu_{k_t(x)}\| > r | k_t(x) = j\right) + \sum_j \pi_j \mathbb{P}\left(\rho_t(X_t) < \rho^* | k_t(x) = j\right)$$

Using Gaussian concentration around its mean,

$$\mathbb{P}\left(\|X_t - t\mu_{k_t(x)}\| > r | k_t(x) = j\right) \leq \exp\left(-\frac{r^2}{2B_t(j)}\right).$$

Also, noting the fact that $\rho_t(x) < \rho^*$ implies $\|x - t\mu_{k_t(x)}\| > (tD_{min} - \rho^*)/2$, and again using Gaussian concentration, we have,

$$\mathbb{P}\left(\rho_t(X_t) < \rho^* | k_t(x) = j\right) \leq \exp\left(-\frac{(tD_{min} - \rho^*)^2}{8B_t(j)}\right).$$

Recalling that $\sum_j \pi_j = 1$, and $B_t^{\min} \triangleq \min_j B_t(j)$. Therefore, we get

$$\mathbb{P}\left(X_t \notin \mathcal{G}_t(r, \rho^*)\right) \leq \exp\left(-\frac{r^2}{2B_t^{\min}}\right) + \exp\left(-\frac{(tD_{min} - \rho^*)^2}{8B_t^{\min}}\right),$$

completing the proof. □

Now, for any $x \in \mathcal{G}_{r,\rho^*}$, let $k = k_t(x)$. For every $j \neq k$,

$$\begin{aligned}
\frac{\mathcal{N}(t\mu_j, B_t(j)I_d)}{\mathcal{N}(t\mu_k, B_t(k)I_d)} &= \left(\frac{B_t(k)}{B_t(j)}\right)^{d/2} \exp\left(-\frac{\|x - t\mu_j\|^2}{2B_t(j)} + \frac{\|x - t\mu_k\|^2}{2B_t(k)}\right) \\
&\overset{\text{(i)}}{\leq} \left(\frac{B_t(k)}{B_t(j)}\right)^{d/2} \exp\left(-\frac{\rho^{*2}}{2B_t(j)} + \frac{r^2}{2B_t(k)}\right),
\end{aligned} \tag{60}$$

where (i) follows from $\|x - t\mu_k\| \le r$, and $\|x - t\mu_j\| \ge \rho^*$ (loose bound from margin). Subsequently, the mixture can be written as:

$$\begin{aligned}
&\pi_k \mathcal{N}(t\mu_k, B_t(k)I_d) + \sum_{j \neq k} \pi_j \mathcal{N}(t\mu_j, B_t(j)I_d) \\
&\qquad = \pi_k \mathcal{N}(t\mu_k, B_t(k)I_d) \left(1 + \sum_{j \neq k} \frac{\pi_j \mathcal{N}(t\mu_j, B_t(j)I_d)}{\pi_k \mathcal{N}(t\mu_k, B_t(k)I_d)} \right) \\
&\qquad \overset{\text{(i)}}{=} \pi_k \mathcal{N}(t\mu_k, B_t(k)I_d)(1 + \zeta_t(x)),
\end{aligned}$$

where, using (60) we have,

$$\zeta_t(x) \le \left(\frac{1 - \pi_k}{\pi_k} \right) \left(\frac{B_t(k)}{B_t(j)} \right)^{d/2} \exp\left(-\frac{\rho^{*2}}{2B_t(j)} + \frac{r^2}{2B_t(k)} \right).$$

Clearly, for large margin ρ^* and small r, $\zeta_t(x)$ goes to zero, and the kth Gaussian component dominates.

A.6 Convergence of FM and TM iterates for a Gaussian mixture target

For all the proofs in the section, let $p_1 = \sum_{j=1}^{K} \pi_j \phi_j$ be the Gaussian mixture density, where $\pi_j > 0$, $\sum_j \pi_j = 1$, and $\phi_j \equiv \mathcal{N}(\mu_j, \sigma_j^2 I_d)$, and denote $D_{\min} = \min_{i \neq j} \|\mu_i - \mu_j\|$. The majority of the proofs for Gaussian mixture will proceed component-wise, i.e., given a sample X, we will consider the likelihood that a particular component of the mixture was used for sampling. For this purpose, we introduced the random variable $Z \in \{1, \dots, K\}$ with $\mathbb{P}(Z = j)$ and conditional $X | Z = j \sim \phi_j$.

We prove our main result (Thm. 2) through a series of Lemmas. We will use the notion of *good region* (58), i.e.,

$$\mathcal{G}_t(r, \rho^*) \triangleq \left\{ x : \|x - t\mu_{k_t(x)}\| \le r \text{ and } \rho_t(x) \ge \rho^* \right\}.$$

Lem. 2 shows that if the FM/TM Euler steps bring any iterate close enough to a specific component, then subsequent iterates will also be close to that component.

Lemma 2. (Local component attraction) *For any $M \in \{0, \dots, N-1\}$ and $\beta \in \left(0, \frac{1}{2}\right)$, set $r = \beta t_M D_{\min}$ and $\rho^* = (1 - 2\beta) t_M D_{\min}$. Fix some $M \in \{0, \dots, N-1\}$, and let $\sigma_{\max} = \max_j \sigma_j$, $B_{\max}(t) = (1-t)^2 + \sigma_{\max}^2$ and $B_* = \max_{m \ge M} B_{\max}(t_m) \le (1 - t_M)^2 + \sigma_{\max}^2$. If at the "hitting time" M, the iterate lies in the good region for some k, i.e.,*

$$\widetilde{X}_{t_M}^A \in \mathcal{G}_{t_M}(r, \rho^*) = \left\{ x : \|x - t_M \mu_{k_{t_M}(x)}\| \le r, \quad \rho_{t_M}(x) \ge \rho^* \right\} \quad \textit{with } k_{t_M}\left(\widetilde{X}_{t_M}^A \right) = k,$$

then for all $m \ge M$, we have,

$$\widetilde{X}_{t_m}^A \in \mathcal{G}_{t_m}(r, \rho^*) \quad \textit{and} \quad k_{t_m}\left(\widetilde{X}_{t_m}^A \right) = k, \quad \textit{with probability exceeding } 1 - \delta,$$

where $\delta = (N - M + 1) \exp\left(-\frac{1}{2} \left(\frac{r}{\sqrt{B_*}} - \sqrt{d} \right)_+^2 \right)$.

Proof. For either algorithm $A \in \{\text{FM}, \text{TM}\}$, conditioning on a specific component label $Z = k$, from the proof of Thm. 3, we have,

$$\mathbb{E}\left[\widetilde{X}_{t_n}^A | Z = k \right] = t_n \mu_k, \quad \text{Cov}\left(\widetilde{X}_{t_n}^A | Z = k \right) = s_{n,k}^A I_d, \tag{61}$$

$$\text{where } s_{n,k}^A \le B_{t_n}(k) = (1 - t_n)^2 + t_n^2 \sigma_k^2.$$

We assume that at the *hitting time* M, the iterate lies in the good region for some k, i.e.,

$$\widetilde{X}_{t_M}^A \in \mathcal{G}_{t_M}(r, \rho^*) \quad \text{and } k_{t_M}(\widetilde{X}_{t_M}^A) = k. \tag{62}$$

Firstly, we show that if we can keep $\widetilde{X}_{t_m}$ inside the Euclidean ball, $\{x : \|x - t_m \mu_k\| \le r\}$ for all future steps $m \ge M$, then it automatically implies that $\widetilde{X}_{t_m} \in \mathcal{G}_{t_m}(r, \rho^*)$ for all $m \ge M$. To see this, note that because t_m is non-decreasing in m and $\beta \in \left(0, \frac{1}{2}\right)$, for every $m \ge M$,

$$r = \beta t_M D_{\min} \le \beta t_m D_{\min}, \quad \text{and } \rho^* = (1 - 2\beta) t_M D_{\min} \le (1 - 2\beta) t_m D_{\min}.$$

Hence, if $\|x - t_m\mu_k\| \le r$, then for any $j \neq k$,

$$\|x - t_m\mu_j\| \ge t_m\|\mu_j - \mu_k\| - \|x - t_m\mu_k\| \ge t_m D_{\min} - \|x - t_m\mu_k\| \overset{\text{(i)}}{\ge} \|x - t_m\mu_k\| + \rho^*, \tag{63}$$

where (i) follows because,

$$2\|x - t_m\mu_k\| + \rho^* \le 2r + \rho^* = 2\beta t_M D_{\min} + (1 - 2\beta)t_M D_{\min} = t_M D_{\min} \le t_m D_{\min}.$$

In other words, (63) implies $\{x : \|x - t_m\mu_k\| \le r\} \subseteq \mathcal{G}_{t_m}(r, \rho^*)$ for all $m \ge M$. Hence, for the remainder of the proof, we only need to show that $\mathbb{P}\left(\|\widetilde{X}_{t_m} - t_m\mu_k\|\right) \le r$ for all $m \ge M$.

Now, note that conditioned on $Z = k$, we have $\widetilde{X}^A_{t_m} - t_m\mu_k \sim \mathcal{N}\left(0, s^A_{m,k} I_d\right)$. Using Gaussian concentration, we obtain the following uniform bound over $m \ge M$:

$$\mathbb{P}\left(\|\widetilde{X}^A_{t_m} - t_m\mu_k\| > r \mid Z = k\right) \le \exp\left(-\frac{1}{2}\left(\frac{r}{\sqrt{B_*}} - \sqrt{d}\right)_+^2\right). \tag{64}$$

An application of union bound over the event,

$$\begin{aligned}\mathcal{E} \triangleq \bigcap_{m=M}^{N} \left\{\|\widetilde{X}^A_{t_m} - t_m\mu_k\| \le r\right\} \implies \mathbb{P}\left(\mathbb{E}^{\mathrm{C}} \mid Z = k\right) \\ \le (N - M + 1)\exp\left(-\frac{1}{2}\left(\frac{r}{\sqrt{B_*}} - \sqrt{d}\right)_+^2\right) \triangleq \delta.\end{aligned}$$

In other words, $\mathbb{P}\left(\forall m \ge M : \widetilde{X}^A_{t_m} \in \mathcal{G}_{t_m}(r, \rho^*) \mid Z = k\right) \ge 1 - \delta$, completing the proof. □

Next, the following Lemma shows that if any x is within the good region for a component k, then with a high probability that component is dominant with respect to the posterior responsibilities.

Lemma 3. *Fix $t \in [0, 1]$. Suppose $x \in \mathcal{G}_t(r, \rho^*)$, then $w_t(x, k) \ge 1 - \epsilon$, where*

$$\epsilon \le \left(\frac{B_t^{\max}}{B_t^{\min}}\right)^{d/2} \left(\frac{1 - \pi_k}{\pi_k}\right) \exp\left(-\frac{(\rho^*)^2}{2B_t^{\max}} + \frac{r^2}{2B_t^{\min}}\right). \tag{65}$$

Proof. Note that,

$$1 - w_t(x, k) = \frac{\sum_{j \neq k} \pi_k \phi_t(j)}{\pi_k \phi_t(k) + \sum_{j \neq k} \pi_j \phi_t(j)} \le \sum_{j \neq k} \frac{\pi_j \phi_t(j)}{\pi_k \phi_t(k)}, \tag{66}$$

where $\phi_t(j) \equiv \mathcal{N}(t\mu_j, B_t(j) I_d)$, and,

$$\begin{aligned}\frac{\phi_t(j)}{\phi_t(k)} &= \left(\frac{B_t(k)}{B_t(j)}\right)^{d/2} \exp\left(-\frac{\|x - t\mu_j\|^2}{2B_t(j)} + \frac{\|x - t\mu_k\|^2}{2B_t(k)}\right) \\ &\overset{\text{(i)}}{\le} \left(\frac{B_t(k)}{B_t(j)}\right)^{d/2} \exp\left(-\frac{(\rho^*)^2}{2B_t^{\max}} + \frac{r^2}{2B_t^{\min}}\right),\end{aligned}$$

where (i) follows because $x \in \mathcal{R}_t(r, \rho^*)$ and $k_t(x) = k$. Summing over all $j \neq k$, eq. (66) simplifies to,

$$1 - w_t(x, k) \le \left(\frac{B_t^{\max}}{B_t^{\min}}\right)^{d/2} \left(\frac{1 - \pi_k}{\pi_k}\right) \exp\left(-\frac{(\rho^*)^2}{2B_t^{\max}} + \frac{r^2}{2B_t^{\min}}\right).$$

This completes the proof. □

A.7 Proof of Thm. 2: Mixture of Gaussians Target Distribution

Theorem 4. *For any $M \in \{0, \dots, N-1\}$ and $\beta \in \left(0, \frac{1}{2}\right)$, set $r = \beta t_M D_{\min}$ and $\rho^* = (1 - 2\beta) t_M D_{\min}$. Suppose $\widetilde{X}_{t_M}^{(\cdot)} \in \mathcal{G}_{t_M}(r, \rho^*)$, where $(\cdot)$ is either FM or TM, and r and ρ^*. Then,*

$$\mathrm{KL}(p_1^{\mathrm{TM}} \| p_1) < \mathrm{KL}(p_1^{\mathrm{FM}} \| p_1) - \gamma,$$

where γ can be arbitrarily close to 0, and $p_1^{(\cdot)}$ are the marginal distributions of the final iterates.

Proof. From Lem. 2, the final iterates for FM and TM satisfy, $\widetilde{X}_1^{(\cdot)} \in \mathcal{G}_1(r, \rho^*)$ with probability exceeding $1 - \delta$. Subsequently, from Lem. 3, with probability exceeding $1 - \delta$, the posterior responsibility of the final iterate satisfies $w_1(\widetilde{X}_1^{(\cdot)}, k) \geq 1 - \epsilon$. From Lem. 6, for $\Delta_{\mathrm{FM}}, \Delta_{\mathrm{TM}} \in [\log(1-\epsilon), 0]$, we have,

$$\mathrm{KL}(p_1^{\mathrm{FM}} \| p_1) = \mathrm{KL}(p_1^{\mathrm{FM}} \| \phi_k) + \Delta_{\mathrm{FM}}, \quad \text{and} \quad \mathrm{KL}(p_1^{\mathrm{TM}} \| p_1) = \mathrm{KL}(p_1^{\mathrm{TM}} \| \phi_k) + \Delta_{\mathrm{TM}}.$$

From the unimodal analysis Thm. 1, we have $\mathrm{KL}(p_1^{\mathrm{TM}} \| \phi_k) < \mathrm{KL}(p_1^{\mathrm{FM}} \| \phi_k)$. This implies,

$$\mathrm{KL}(p_1^{\mathrm{FM}} \| p_1) - \mathrm{KL}(p_1^{\mathrm{TM}} \| p_1) > \Delta_{\mathrm{FM}} - \Delta_{TM} \triangleq \gamma. \tag{67}$$

Since Δ_{FM} and Δ_{TM} can be made arbitrarily close to 0, γ is also arbitrarily close to 0, completing the proof. □

B Preliminaries: Probability Theory

This section states some useful results from probability theory that are useful for proving the main results in App. A. Most of them can be easily proved.

Lemma 4. (Conditional distribution of jointly Gaussian vectors) *Let (Y, X) be jointly Gaussian-distributed, and let its joint density in block-partitioned format be represented as*

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}, \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{bmatrix} \right).$$

Then, the conditional distribution $Y|X = x$ is also Gaussian with mean and covariance given by,

$$\mathbb{E}[Y|X = x] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X) \quad \textit{and,} \quad \mathrm{Cov}(Y|X = x) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

Proof. Consider the inverse Covariance matrix $J \triangleq \Sigma^{-1}$ in block format as

$$J \;=\; \begin{bmatrix} J_{YY} & J_{YX} \\ J_{XY} & J_{XX} \end{bmatrix}.$$

Then, the joint Gaussian density can be written as,

$$p(y, x) \;\propto\; \exp\left(-\tfrac{1}{2} \begin{bmatrix} y - \mu_Y \\ x - \mu_X \end{bmatrix}^\top J \begin{bmatrix} y - \mu_Y \\ x - \mu_X \end{bmatrix} \right).$$

When x is fixed, the terms depending on y are

$$-\tfrac{1}{2}(y - \mu_Y)^\top J_{YY}(y - \mu_Y) \;-\; (y - \mu_Y)^\top J_{YX}(x - \mu_X) \;+\; \mathrm{const}(x).$$

Completing the square in y yields,

$$-\tfrac{1}{2}\left(y - \mu_Y + J_{YY}^{-1}J_{YX}(x - \mu_X)\right)^\top J_{YY}\left(y - \mu_Y + J_{YY}^{-1}J_{YX}(x - \mu_X)\right) \;+\; \mathrm{const}(x).$$

Therefore,

$$Y \mid X = x \;\sim\; \mathcal{N}\big(\mu_Y - J_{YY}^{-1}J_{YX}(x - \mu_X),\; J_{YY}^{-1}\big).$$

Moreover, using standard block inverse identities,

$$J_{YY}^{-1} \;=\; \Sigma_{YY} - \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY}, \qquad J_{YY}^{-1}J_{YX} \;=\; -\,\Sigma_{YX}\,\Sigma_{XX}^{-1}.$$

Substituting these expressions above completes the proof. □

Lemma 5. (KL divergence between multivariate Gaussian distributions) *Let* $\mathrm{P} \equiv \mathcal{N}(\mu_p, \Sigma_p)$ *and* $\mathrm{Q} \equiv \mathcal{N}(\mu_q, \Sigma_q)$*. Then, the KL divergence between them is given by*

$$\mathrm{KL}\left(\mathrm{P} \parallel \mathrm{Q}\right) = \frac{1}{2}\left[\mathrm{Tr}\left(\Sigma_q^{-1}\Sigma_p\right) + (\mu_q - \mu_p)^\top \Sigma_q^{-1}(\mu_q - \mu_p) - d + \log\frac{\det \Sigma_q}{\det \Sigma_p}\right].$$

Lemma 6. (KL divergence of mixture and dominant component) *Let* $p_1 = \sum_{j=1}^{K} \pi_j \phi_j$*, where* $\pi_j > 0$*,* $\sum_j \pi_j = 1$*, and* $\phi_j \equiv \mathcal{N}(\mu_j, \sigma_j^2 I_d)$ *be the Gaussian mixture density. Fix a component index* $k \in \{1, \ldots, K\}$*. For* x *with* $p_1(x) > 0$*, denote*

$$w(x, k) = \frac{\pi_k \phi_k(x)}{p_1(x)} \in (0, 1]. \tag{68}$$

For any probability distribution q *on* $\mathbb{R}^d$*, assume that for some arbitrarily small* $\epsilon \in [0, 1)$*,* $w(x, k) \geq 1 - \epsilon$ *for* q*-almost every* x*. In other words, the* k^{th} *component is dominant. Then,*

$$\mathrm{KL}(q\|p_1) = \mathrm{KL}(q\|\phi_k) + \log\frac{1}{\pi_k} + \Delta, \quad \Delta \in [\log(1-\epsilon), 0]. \tag{69}$$

In order words, the additive gap from replacing the mixture p_1 *by a single component* ϕ_k *is bounded by*

$$\left|\mathrm{KL}(q\|p_1) - \left(\mathrm{KL}(q\|\phi_k) + \log\frac{1}{\pi_k}\right)\right| \leq -\log(1-\epsilon). \tag{70}$$

Proof. Note that $w(x, k)$ denotes the posterior probability that x was drawn from the kth component of the mixture p_1. Using (68), we have,

$$\begin{aligned}
\log p_1(x) &= \log \pi_k + \log \phi_k(x) - \log w(x, k) \\
\implies \mathrm{KL}(q\|p_1) &\overset{\text{(i)}}{=} \mathbb{E}_{X\sim q}\left[\log q(X) - \log p_1(X)\right] \\
&= \mathbb{E}_{X\sim q}\left[\log q(X) - \log \phi_k(X)\right] - \log \pi_k + \mathbb{E}_{X\sim q}[\log w(X, k)] \\
&= \mathrm{KL}(q\|\phi_k) + \log\frac{1}{\pi_k} + \Delta,
\end{aligned} \tag{71}$$

where (i) follows from the definition of KL divergence, and Δ denotes $\mathbb{E}_{X\sim q}[\log w(X, k)] \in [\log(1-\epsilon), 0]$. The bounds on Δ are obtained by noting that $1 - \epsilon \leq w(x, k) \leq 1$, completing the proof. □

C Related Work

Diffusion and Flow Models. Diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021b) synthesize data by inverting a forward process that progressively injects Gaussian noise. Training regresses targets implied by this corruption; the standard choice is noise prediction Ho et al. (2020); Song et al. (2021a), with clean-data prediction Kingma et al. (2021) and v-prediction Salimans & Ho (2022) which balances the previous choices. Flow models Lipman et al. (2023); Liu et al. (2023); Albergo et al. (2023) instead consider a continuous-time path from a source to the target distribution and learn the instantaneous velocity field that transports samples along that path. Despite their different parameterization, both families learn conditional transition kernels determined by a supervising forward process Holderrieth et al. (2024); Gao et al. (2024). Transition Matching Shaul et al. (2025) generalizes this perspective and can naturally be applied to different choices of parameterization.

Expectation and Conditional Velocity. Karras et al. (2022); Xu et al. (2023); Biroli et al. (2024) proposed and analyzed exact score function of an ideal denoiser in diffusion models. Scarvelis et al. (2023) leveraged exact score to investigate the generalization ability in diffusion models. In the context of flow models, Ryzhakov et al. (2024) proposed to leverage the expectation velocity in closed-form to reduce the variance during training. Close to our work, Bertrand et al. (2025) analyzed expectation and conditional velocity, but focused on the identification on the sources of generalization in flow models. Previous work Bertrand et al. (2025) also presented analysis on the velocity distribution using `Two Moons` and CIFAR-10 datasets where the effects of data separation and variance are entangled. In this work, we present systemic analysis on controlled Gaussian settings to disentangle the effects of mean separation and target variance.

Table 1: **Number of Model Parameters and Computation Costs of TM and FM.** Number of model parameters is reported in millions, and computation time is reported in seconds.

	Backbone	Head	Ratio	C_B	C_H	κ
Class-Conditioned Image Generation						
FM	231.15	-	-	0.00979	-	-
TM	204.49	36.00	6.68	0.01120	0.00238	4.70
Frame-Conditioned Video Generation						
FM	625.85	-	-	0.00962	-	-
TM	627.08	9.08	69.06	0.00965	0.00024	40.08

D Implementation Details

Class-Conditioned Image Generation. For image generation task, we train both FM and TM on ImageNet-10K (Deng et al., 2009) in a class-conditional setting at 256×256 resolution. The images are encoded into latents using KL-16 (Rombach et al., 2022), and training is carried out for 400 epochs for both models.

FM is implemented following the original DiT work (Peebles & Xie, 2023), and TM is implemented from scratch following its original work (Shaul et al., 2025). The TM flow head is a 6-layer MLP, with the backbone latent injected via AdaLN (Peebles & Xie, 2023). Following the prior work MAR (Li et al., 2024), the images are encoded using a pretrained VAE (Rombach et al., 2022) into a latent representation of size $16 \times 16 \times 16$.

Frame-Conditioned Video Generation. For video generation, we adopt the History-Guided Diffusion framework (Song et al., 2025), which utilizes 3D DiT blocks. HGD is a variant of FM that only differs by the interpolation path[2]. For notational conciseness, we therefore refer to HGD as FM.

Following (Song et al., 2025), we train both FM and TM for $360K$ steps on Kinect-600 (Shaul et al., 2023) at 128×128 resolution, where videos are encoded using a pretrained VAE. Each video frame is encoded by a pretrained VAE into a latent tensor of shape $16 \times 16 \times 16$, and the temporal dimension is downsampled by a factor of 4. Given the latent frames, we apply 3D RoPE positional embeddings (Su et al., 2024) for positional encoding. Similar to the image generation task, the flow head is a 6-layer MLP with hidden dimension 384. At test time, the model conditions on a single frame and generates four future frames. For all experiments, we adopt DiT (Peebles & Xie, 2023) for backbone models, and the flow head of TM is implemented using MLP layers.

The number of model parameters and computation costs for each module are reported in Tab. 1. The backbone model size and computation cost C_B are kept comparable for both FM and TM, while the TM flow head remains relatively small and computationally efficient, with cost C_H. Hence, the cost ratio between C_B and C_H, κ, shows a several-fold difference, indicating that increasing S can be more efficient than increasing N, particularly for small N, as shown in (12). For all other model hyperparameters, we follow the baseline configurations, and all models are trained using 32 NVIDIA A100 GPUs.

E Additional Results

Additional Results on Unimodal Gaussian. In this section, we extend the experiments from §3 and present a quantitative comparison between FM and TM under varying compute budget N with fixed S. For TM, the inner ODE steps are fixed to $S \in \{1, 2\}$, and N is chosen respectively for FM and TM to ensure comparable compute (wall clock time). For each task, we adopt the same evaluation setup as described in the main paper.

The results are shown in Fig. 6. While both FM and TM converge to zero KL divergence with increased compute (larger wall clock time), TM consistently outperforms FM under a fixed compute budget (red dotted line), particularly in the low-step sampling regime. Note that TM with a larger

[2]HGD utilizes Variance Preserving (Ho et al., 2020) interpolation path.

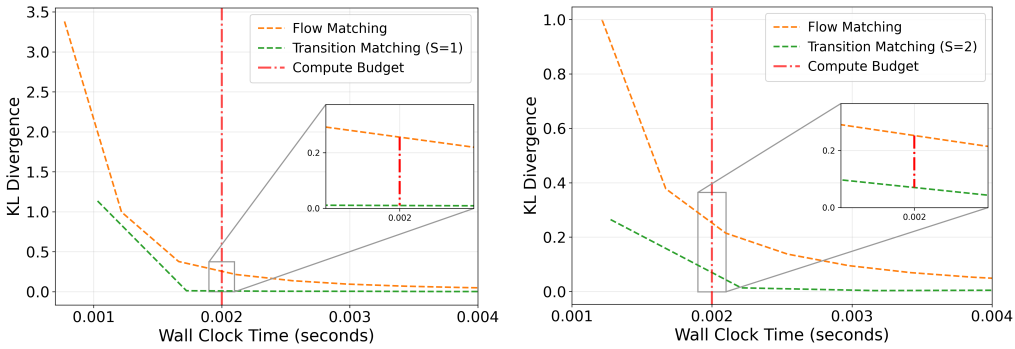

Figure 6: **Unimodal Gaussian KL Divergence against Wall Clock Time.** We compare Flow Matching (FM) and Transition Matching (TM) on unimodal Gaussian KL divergence against wall clock time. For TM, we fix $S \in \{1, 2\}$. The inset zooms a reference region: at matched wall-clock compute, TM achieves lower KL than FM in the low-step regime.

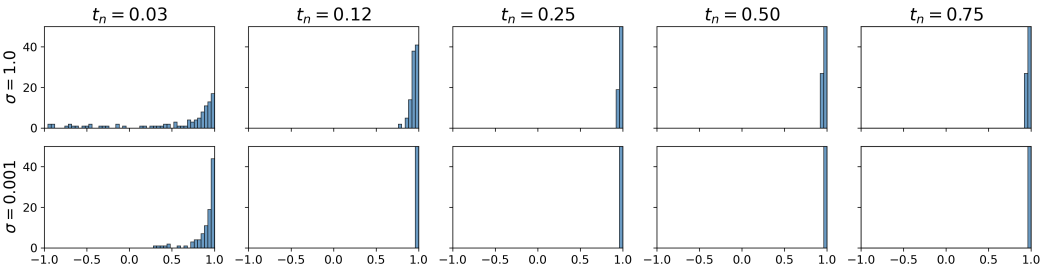

Figure 7: **Effect of Target Variance on $p(V|X)$.** High-dimensional mixture of Gaussian with two variance settings, $\sigma \in \{1.0, 0.001\}$. Each panel shows the histogram of cosine similarities between difference latent samples $\widetilde{V}_{t_n} \sim p(V \mid X_{t_n})$ and its expectation $\mathbb{E}[V \mid X_{t_n}]$ at timestep t_n. For $\sigma = 1.0$, the difference latent samples form a unimodal distribution with non-negligible variance, whereas for $\sigma = 0.001$, they concentrate near 1, indicating collapse toward the mean.

number of inner ODE steps ($S = 2$) yields lower KL divergence, reflecting the result predicted by Cor. 1 of the main paper.

Additional Results on the Effect of Target Variance. Extending the experiment presented in §3, we analyze the effect of target variance on difference latent distribution. Here, the target Gaussian mixture is 64-dimensional with $D_{\text{min}} = 45$. As shown in Fig. 7 (row 1), when $\sigma = 1.0$, the cosine similarities cluster near 1 with non-negligible variance, indicating an approximation to a unimodal distribution with non-zero variance. Hence, following Thm. 1, in this regime with non-negligible variance, stochastic latent updates in TM preserve covariance and can outperform FM. As discussed in §3, when the variance approaches zero, the difference latent distribution approximates Dirac distribution (17). Fig. 7 (row 2) shows this collapse: the cosine similarity is sharply concentrated near 1, indicating that the sampled difference latent is nearly identical to its expectation. Since $\mathbb{E}[V \mid X_{t_n}]$ coincides with the velocity used in the FM update (5), the distinction between FM and TM vanishes.

Comparison to Diffusion Models. We compare TM with three distinct sampling strategies: DDPM (Ancestral Sampling) Ho et al. (2020), DDIM (Skip-Step Sampler) Song et al. (2021a), and Optimal Covariance Matching Ou et al. (2024). Given a total number of steps N, the interval $[0, 1]$ is discretized as $t_n = n\Delta t$, where $\Delta t = 1/N$ and $n \triangleq \{0, \ldots, N-1\}$. Following the standard diffusion-model convention, we denote $t_N = 1$ as the initial (noisy) timestep and $t_0 = 0$ as the terminal timestep, opposite to the convention adopted in our paper. The source is the standard Gaussian $X_{t_{N=1}} \sim \mathcal{N}(0, I_d)$, and the target is $X_{t_{0=0}} \sim \mathcal{N}(\mu, \sigma^2 I_d)$.

DDPM (Ancestral Sampling) Ho et al. (2020): Diffusion model defines Markovian forward process with a predefined variance schedule $\beta_{t_n} = \beta_{\min} + t_n(\beta_{\max} - \beta_{\min})$:

$$p\left(X_{t_{n+1}}^{\mathrm{DM}} \mid X_{t_n}^{\mathrm{DM}}\right) = \mathcal{N}\left(\sqrt{1-\beta_{t_n}} X_{t_n}^{\mathrm{DM}}, \beta_{t_n} I_d\right),$$

where the $(\beta_{\min}, \beta_{\max})$ are chosen so that as $n \to N$, $p_{t_N} \to \mathcal{N}(0, I_d)$, e.g., $N = 1000$. Under this choice of noise scheduling, DDPM samples follow a *curved* probability path. The corresponding single-step reverse process is governed by the transition kernel:

$$p\left(\widetilde{X}_{t_{n-1}}^{\mathrm{DM}} \mid \widetilde{X}_{t_n}^{\mathrm{DM}}\right) = \mathcal{N}\left(\frac{\widetilde{X}_{t_n}^{\mathrm{DM}} + \beta_{t_n} \nabla_x \log p_{t_n}\left(\widetilde{X}_{t_n}^{\mathrm{DM}}\right)}{\sqrt{1-\beta_{t_n}}}, \beta_{t_n} I_d\right),$$

where the variance is fixed *heuristically* to β_{t_n}. However, ancestral sampling in DDPM requires a fine-grained time discretization (large N) in order to accurately approximate the target distribution.

DDIM (Skip-Step Sampler) Song et al. (2021a): DDIM generalizes DDPM, allowing for skip-step sampling. Denote $\alpha_{t_n} = 1 - \beta_{t_n}$, $\bar{\alpha}_{t_{n-1}:t_n} = \prod_{s=t_{n-1}}^{t_n}(\alpha_s)$, and $\bar{\alpha}_{t_n} = \prod_{s=1}^{t_n}(\alpha_s)$. Then DDIM sampler utilizes the following reverse transition kernel:

$$p\left(\widetilde{X}_{t_{n-1}}^{\mathrm{DI}} | \widetilde{X}_{t_n}^{\mathrm{DI}}\right) = \mathcal{N}\left(\frac{\widetilde{X}_{t_n}^{\mathrm{DI}} + (1-\bar{\alpha}_{t_{n-1}:t_n})\nabla_{x_{t_n}} \log p_\theta\left(\widetilde{X}_{t_n}^{\mathrm{DI}}\right)}{\sqrt{\bar{\alpha}_{t_{n-1}:t_n}}}, (\sigma_{t_n}^{\mathrm{DI}})^2 I_d\right),$$

where the variance follows the heuristic choice of DDPM, $\sigma_{t_n}^{\mathrm{DI}} = \sqrt{(1-\bar{\alpha}_{t_{n-1}})/(1-\bar{\alpha}_{t_n})}\sqrt{1-\bar{\alpha}_{t_{n-1}:t_n}}$.

Optimal Covariance Matching (OCM) Ou et al. (2024): OCM addresses the limitation of heuristic variances by explicitly modeling the optimal covariance of the posterior $p(x_{t_{n-1}}|x_{t_n})$:

$$\Sigma_{t_{n-1}|t_n}^{\mathrm{opt}}(x) = ((1-\bar{\alpha}_{t_n})^2 \nabla_x^2 \log p_{t_n}(x) + (1-\bar{\alpha}_{t_n}) I_d)/\bar{\alpha}_{t_n},$$

which can be used to replace the heuristic covariance of the previous samplers. In practice, the Hessian is intractable, and OCM uses an additional network that learns an approximation of this quantity. Using the optimal covariance gives the following transition kernel:

$$p\left(\widetilde{X}_{t_{n-1}}^{\mathrm{OCM}} | \widetilde{X}_{t_n}^{\mathrm{OCM}}\right) = \mathcal{N}\left(\frac{\widetilde{X}_{t_n}^{\mathrm{OCM}} + (1-\bar{\alpha}_{t_{n-1}:t_n})\nabla_{x_{t_n}} \log p_\theta\left(\widetilde{X}_{t_n}^{\mathrm{OCM}}\right)}{\sqrt{\bar{\alpha}_{t_{n-1}:t_n}}}, \Sigma_{t_{n-1}|t_n}^{\mathrm{opt}}(\widetilde{X}_{t_n}^{\mathrm{OCM}})\right).$$

Empirical Evaluation. In Tab. 2, we compare the KL divergence of these samplers on a unimodal Gaussian target. For DDPM, DDIM, OCM, and FM, we scale the number of integration steps N. For TM, we fix $N = 1$ and scale S. Note that this regime affords TM a lower total compute budget than the baselines at each case. Additionally, for OCM, we utilize the ground truth optimal covariance which can be derived in closed form for this unimodal Gaussian case.

Table 2: **Quantitative Comparison of FM and TM to Diffusion Models.**

	Number of Steps					
Model	2	4	8	16	128	1000
DDPM Ho et al. (2020)	73.81	72.89	71.00	68.30	36.38	0.00
DDIM Song et al. (2021a)	6.88	1.42	0.40	0.15	0.01	0.00
OCM Ou et al. (2024)	7.48	0.01	0.00	0.00	0.00	0.00
FM Gat et al. (2024)	0.92	0.21	0.06	0.02	0.00	0.00
TM ($N = 1$) Shaul et al. (2025)	0.28	0.11	0.05	0.04	0.01	0.00

As expected, classic DDPM ancestral sampling incurs significant errors when skipping steps, as its transition kernel is derived under the assumption of Markov process. While DDIM mitigates this by reformulating the reverse process, it still exhibits non-negligible divergence in the few-step regime ($N \leq 8$); this stems from its heuristic variance choice, which fails to capture the true conditional covariance of the strided transition. In contrast, OCM achieves near-perfect convergence even at very

few steps ($N = 4$) by leveraging the ground truth optimal covariance. However, it is important to note that OCM utilized the ground truth Hessian and covariance for Gaussian targets. In practical scenarios, the ground truth covariance is intractable, forcing OCM to rely on approximations (e.g., via the Hutchinson estimator). On the other hand, TM relaxes the explicit Gaussian assumption and trains a continuous flow to directly model the transition kernel.

Comparison of FM and TM in Anisotropic Gaussian. Let $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \mathcal{N}(\mu, \Sigma)$ be independent Gaussian vectors in $\mathbb{R}^d$, where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal covariance. Note that since the source covariance I_d and the target covariance Σ are both diagonal, and $X_0 \perp X_1$, the components of the vector $X_t = (1-t)X_0 + tX_1$ are mutually independent for all $t \in [0, 1]$. Consequently, the d-dimensional problem decouples, and the proof follows a similar logic to the isotropic case shown in Thm. 1.

Empirical Evaluation. In Tab. 3, we evaluate FM and TM on a 2D anisotropic Gaussian target with diagonal covariance $\Sigma = \text{diag}(0.5, 1.5)$. We compare FM (scaling N) against TM (scaling S with $N = 1$). As shown in the table, TM yields lower KL divergence than FM in the few-step regime, confirming its superior efficiency in Gaussian target with anisotropic covariance.

Table 3: **Quantitative Comparison of FM and TM in Anisotropic Gaussian.**

Model	**2**	**4**	**8**	**16**	**32**
FM Lipman et al. (2023)	0.80	0.20	0.06	0.02	0.01
TM Shaul et al. (2025) ($N = 1$)	0.13	0.06	0.05	0.03	0.01

Additional Results with Varying N and S. We present additional quantitative results comparing FM and TM with varying the number of outer steps (N) and inner ODE steps (S). For image generation (Tab. 4-5), TM outperforms FM in both quality and efficiency. Specifically, TM ($N = 8, S = 8$) achieves a higher IS and lower FID than FM ($N = 32$), while reducing inference latency. In the video generation task (Tab. 6-7), we observe a similar trend: FM ($N = 16$) reaches higher FVD and FID than TM ($N = 12, S = 16$). Beyond scaling N, TM yields performance gains when scaling the inner steps S with a fixed N, contributing to practical computational efficiency.

Table 4: **Quantitative Results of FM in Class-Conditioned Image Generation**.

Metric	N=8	N=16	N=32	N=64
IS ↑	$411.70^{\pm 1.07}$	$432.47^{\pm 1.04}$	$438.96^{\pm 1.56}$	$441.11^{\pm 0.69}$
FID ↓	$23.73^{\pm 0.08}$	$22.13^{\pm 0.05}$	$22.11^{\pm 0.09}$	$21.97^{\pm 0.14}$
Time (s) ↓	0.04	0.08	0.17	0.32

Table 5: **Quantitative Results of TM in Class-Conditioned Image Generation**.

Metric	Step (S)	N=8	N=16	N=32	N=64
IS ↑	2	$436.96^{\pm 0.51}$	$441.49^{\pm 1.13}$	$445.05^{\pm 1.71}$	$445.54^{\pm 1.06}$
	4	$438.39^{\pm 0.88}$	$446.76^{\pm 0.97}$	$448.63^{\pm 1.07}$	$448.09^{\pm 0.77}$
	8	$440.90^{\pm 0.80}$	$449.43^{\pm 0.84}$	$450.58^{\pm 0.88}$	$449.74^{\pm 0.52}$
	12	$441.32^{\pm 0.66}$	$450.22^{\pm 0.91}$	$451.39^{\pm 0.84}$	$449.80^{\pm 1.08}$
	16	$441.19^{\pm 0.62}$	$450.69^{\pm 0.85}$	$451.80^{\pm 0.93}$	$450.14^{\pm 1.27}$
	20	$441.11^{\pm 0.76}$	$450.82^{\pm 0.86}$	$452.13^{\pm 0.92}$	$450.27^{\pm 1.22}$
FID ↓	2	$22.06^{\pm 0.11}$	$22.07^{\pm 0.12}$	$21.49^{\pm 0.06}$	$21.25^{\pm 0.06}$
	4	$21.86^{\pm 0.10}$	$21.79^{\pm 0.07}$	$21.35^{\pm 0.08}$	$21.05^{\pm 0.11}$
	8	$21.46^{\pm 0.08}$	$21.67^{\pm 0.11}$	$21.29^{\pm 0.08}$	$20.99^{\pm 0.10}$
	12	$21.23^{\pm 0.09}$	$21.58^{\pm 0.10}$	$21.27^{\pm 0.08}$	$20.97^{\pm 0.04}$
	16	$21.11^{\pm 0.08}$	$21.54^{\pm 0.09}$	$21.25^{\pm 0.08}$	$20.96^{\pm 0.06}$
	20	$21.03^{\pm 0.08}$	$21.52^{\pm 0.08}$	$21.24^{\pm 0.07}$	$20.95^{\pm 0.06}$
Time (s) ↓	2	0.08	0.16	0.31	0.62
	4	0.10	0.20	0.39	0.78
	8	0.14	0.28	0.55	1.10
	12	0.18	0.35	0.71	1.41
	16	0.22	0.43	0.87	1.73
	20	0.26	0.51	1.02	2.05

Table 6: **Quantitative Results of FM in Frame-Conditioned Video Generation**.

Metric	N=8	N=12	N=16	N=20
FVD ↓	$113.01^{\pm 0.23}$	$95.06^{\pm 0.15}$	$83.69^{\pm 0.12}$	$76.40^{\pm 0.08}$
FID ↓	$2.65^{\pm 0.03}$	$1.95^{\pm 0.04}$	$1.73^{\pm 0.01}$	$1.65^{\pm 0.01}$
Time (s) ↓	0.04	0.06	0.08	0.10

Table 7: **Quantitative Results of TM in Frame-Conditioned Video Generation**.

Metric	Step (S)	N=8	N=12	N=16	N=20
FVD ↓	2	$134.13^{\pm 0.19}$	$121.00^{\pm 0.29}$	$112.15^{\pm 0.18}$	$105.00^{\pm 0.29}$
	4	$108.41^{\pm 0.31}$	$100.14^{\pm 0.40}$	$94.39^{\pm 0.30}$	$89.43^{\pm 0.44}$
	8	$92.95^{\pm 0.25}$	$84.60^{\pm 0.38}$	$79.60^{\pm 0.26}$	$76.56^{\pm 0.22}$
	12	$89.28^{\pm 0.24}$	$79.73^{\pm 0.31}$	$74.41^{\pm 0.26}$	$71.38^{\pm 0.39}$
	16	$88.11^{\pm 0.23}$	$77.61^{\pm 0.27}$	$71.94^{\pm 0.22}$	$68.88^{\pm 0.19}$
	20	$87.69^{\pm 0.22}$	$76.48^{\pm 0.26}$	$70.56^{\pm 0.20}$	$67.44^{\pm 0.26}$
FID ↓	2	$1.82^{\pm 0.02}$	$1.51^{\pm 0.01}$	$1.40^{\pm 0.02}$	$1.35^{\pm 0.02}$
	4	$1.66^{\pm 0.03}$	$1.40^{\pm 0.01}$	$1.33^{\pm 0.02}$	$1.30^{\pm 0.03}$
	8	$1.72^{\pm 0.03}$	$1.42^{\pm 0.02}$	$1.34^{\pm 0.03}$	$1.31^{\pm 0.03}$
	12	$1.80^{\pm 0.04}$	$1.46^{\pm 0.02}$	$1.36^{\pm 0.03}$	$1.33^{\pm 0.02}$
	16	$1.86^{\pm 0.04}$	$1.49^{\pm 0.02}$	$1.38^{\pm 0.02}$	$1.35^{\pm 0.03}$
	20	$1.90^{\pm 0.05}$	$1.51^{\pm 0.02}$	$1.39^{\pm 0.02}$	$1.36^{\pm 0.02}$
Time (s) ↓	2	0.04	0.06	0.08	0.11
	4	0.04	0.07	0.09	0.11
	8	0.05	0.07	0.10	0.12
	12	0.05	0.08	0.10	0.13
	16	0.05	0.08	0.11	0.14
	20	0.06	0.09	0.12	0.15

Class-Conditioned Image Generation Qualitative Results. Qualitative results of class-conditioned image generation in §5.2 are presented in Fig. 8. For each image, we use the same class label for both FM and TM. Additionally, we set $N = 64$ for FM and $N = 16$, $S = 8$ for TM. Note that in this setting, TM requires less compute (latency) than FM, as reported in §5.2. Despite the lower compute, TM (bottom) generates images with intricate details, such as the decorations of the throne and the lighting of the fountain.

Frame-Conditioned Video Generation Qualitative Results. Qualitative results of frame-conditioned video generation in §5.3 are presented in Fig. 9. The leftmost column shows the conditioning frame used for video generation, and the two groups on the right show three frames generated by FM and TM, respectively. Videos produced by FM often exhibit artifacts, including missing content from the conditioning frame (e.g., the presenter hand in row 1 and the right leg of the baby in row 2), whereas TM preserves previous frame content. Additionally, TM is also more robust to semantic drift where FM often fails (e.g., the wall in row 3 and the person's face in row 4). Finally, FM frequently yields near-stationary motion (e.g., the legs in row 4).

(a) FM Lipman et al. (2023)

(b) TM Shaul et al. (2025)

Figure 8: **Class-Conditioned Image Generation Qualitative Results.**

Figure 9: **Frame-Conditioned Video Generation Results.**