Flow Matching for Multimodal Distributions

Gaoxiang Luo*[1] Frank Cole*[2] Sihang Zhang[3] Yuxiang Wan[1] Yulong Lu[2] Ju Sun[1]
[1]Department of Computer Science and Engineering [2]School of Mathematics [3]School of Statistics
University of Minnesota Twin Cities *Equal Contribution
https://mm-flow.github.io

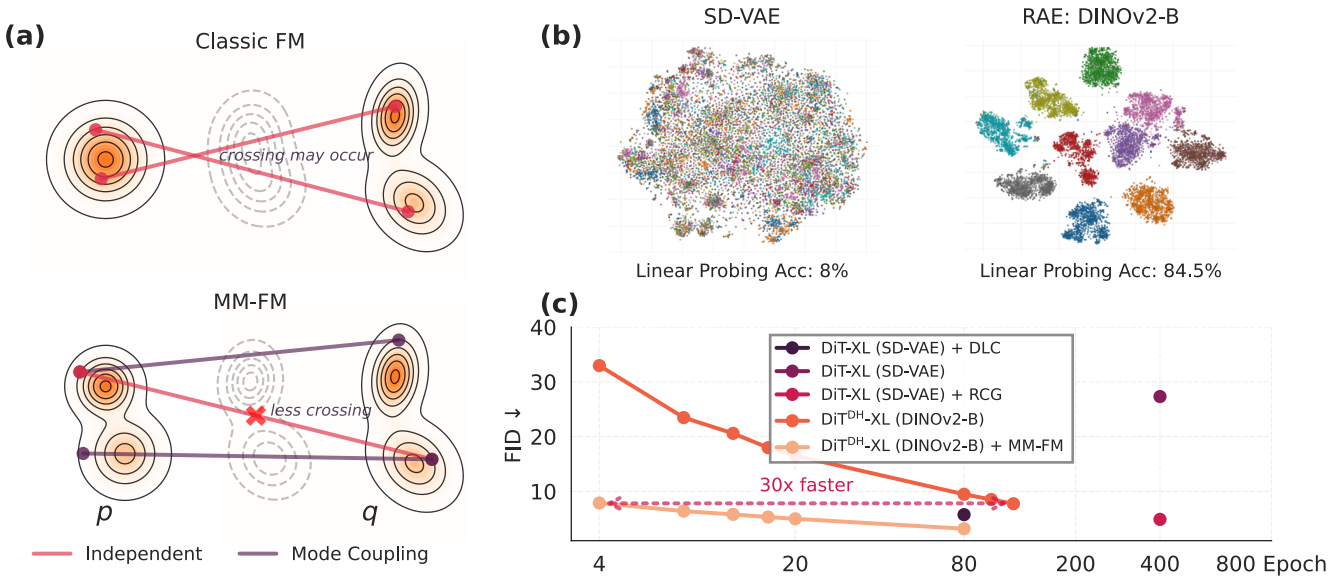

Figure 1. (a) The intuition of source and coupling co-design is to reduce crossing (b) when the target distribution exhibits multimodal structures. (c) Our Multimodal Flow Matching (MM−FM) brings 30× faster convergence for DiT^{DH}-XL trained on DINOv2-B latent space for unconditional generation on ImageNet256 compared with the classic flow matching algorithm.

Abstract

Recently, vision foundation models have been shown to boost the efficiency of flow-based generative models by revealing the intrinsic union-of-manifold structures and lowering the complexity of the latent/target distribution. In this paper, we exploit the multimodality aspect of the union-of-manifold structures, and aim to further improve the learning and inference efficiency for flow-matching models. To this end, we propose an efficient source and coupling co-design method termed Mixture-Modeling Flow Matching (MM−FM), by integrating a data-adaptive multimodal source distribution (implemented as Gaussian mixture models) and mode-dependent data coupling. The former shortens the distance between the source and the target, and the latter promotes local and straighter flows. We also derive theoretical results to confirm our intuition in a quantitative sense. In our experiments on ImageNet256x256 with multimodal DINOv2-B latents, MM−FM exhibits superior learning efficiency and state-of-the-art unconditional generation quality: FID=2.74 with autoguidance in only 80 epochs.

1. Introduction

Most state-of-the-art (SOTA) deep generative models are flow-based [1, 7, 20, 81], predominately diffusion [46] and flow matching (FM) models [53]. In flow-based models, given a target distribution represented by a training dataset, a chosen source distribution is gradually transformed into the target following properly designed flows that move around the probability mass (or samples). Once such flows are learned, they are used to produce novel target samples by transforming new source samples.

Intuitively, the level of learning difficulty incurred by flow-based models is dictated by several factors, including at least **(1)** the complexity of the target distribution; **(2)** the distance between the source and target distribu-

tions [22, 23]; and **(3)** the flow design [15, 80]. Since flow-based models are typically trained in latent spaces, for any given training set, **(1)** is determined by the latent embedding method, i.e., the tokenizer. Related to this, a series of recent works have shown that foundation models can improve variational-autoencoder (VAE)-based tokenizers in terms of the subsequent training efficiency and generation quality [6, 10, 11, 24, 70, 85, 89]. This is partially due to the lower complexity of the resulting *multimodal distributions*[1] [10, 11, 24, 70, 89]—consistent with the *Union of Manifolds Hypothesis* [3, 8, 65] for image data, than that of the entangled distributions induced by VAE-based tokenizers alone (see Fig. 1 (b)). Given that foundation models lead to such low-complexity multimodal target distributions, this paper focuses on a natural question: *are there **principled** choices of the source distribution and the flow design to make learning even more efficient*?

For this, it seems favorable to make the source and target distributions sufficiently close. The closeness not only means small transport distances but also implies that only localized and minuscule flows are needed to move the probability mass around to bridge the two distributions. This consideration rejects the popular choice of unimodal Gaussian as the source [7, 20, 81], which is target-blind.

How can one obtain sufficiently close source distributions, especially for high-dimensional data, in practice? In this paper, we propose using Gaussian Mixture Models (GMM) fitted to the training samples as data-adaptive warm-start source distributions. There are a couple of advantages to this choice: (1) the multiple modes in the GMM model the multimodality aspect of the target distribution; (2) more importantly, the natural assignment of training samples to individual mixture modes after GMM estimation enables us to design mode-dependent data coupling and, hence, flows. By pairing each training sample with source samples from its nearest mode, we can ensure that probability mass moves locally rather than crossing distant modes, significantly cutting down the overall flow complexity.

Our contributions include: (1) **proposing** a source and coupling co-design algorithm for flow-based models that fully exploits the multimodality nature of target distributions (e.g., when foundation models are used as tokenizers), leading to improved training efficiency (faster convergence with better generation quality), inference efficiency (fewer sampling steps), and data efficiency (less required training data) (see Sec. 3.1); (2) **deriving** theoretical insights into the sampling trajectory complexity (e.g., straightness and length) and the learning complexity (e.g., Lipschitz constant) to explain why the proposed modifications can boost efficiency (see Sec. 3.4); (3) **designing** a subroutine to operationally perform GMM estimation for the source distribution and subsequent mode coupling, despite the high data dimensionality; and (4) **conducting** systematic experiments on ImageNet256 dataset to demonstrate 30× faster training convergence, 5× faster inference measured by required ODE steps, and 3× lower terminal Fréchet Inception Distance (FID) when trained on only 10% of the data, compared with the classic flow-matching models (see Sec. 4).

[1]The term "multimodal distribution" is standard in statistics, referring to a distribution with multiple local maxima (or *modes*) in its probability density function. This should not be confused with "multimodal foundation models" that are increasingly used nowadays, where "multimodal" refers to multiple data types and modalities.

2. Technical Background and Related Work

2.1. Flow-based Generative Models

Continuous normalizing flow. At a high level, continuous normalizing flows (CNFs) generalize classic normalizing flows, which are discrete in time. CNFs sample from a target distribution q by *continuously* transforming samples from a source distribution p [14, 26]. The entire transformation path is governed by an invertible continuous-time flow function $\psi_t(x) \doteq \psi(t,x) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, where $t \in [0,1]$ indexes time and $x \in \mathbb{R}^d$ is the initial point, so that $\psi_0(x) = x$ and $\psi_1(x)$ is the end point. The flow can be expressed as the solution to an ordinary differential equation (ODE), driven by a velocity field $u_t(x)$:

$$\mathrm{d}\psi_t(x) = u_t\left(\psi_t(x)\right)\mathrm{d}t. \tag{2.1}$$

Consider the random variable $X_0 \sim p$ and write $X_t \doteq \psi_t(X_0) \sim p_t$. The probability path $\{p_t\}_{t=0}^1$ can be computed via the famous change-of-variable formula $p_t(x) = p_0\left(\psi_t^{-1}(x)\right)\left|\det \partial_x \psi_t^{-1}(x)\right|$. Since the formula allows for a closed-form representation of $p_1(x)$, initial efforts in CNFs parameterize $u_t(\boldsymbol{x})$ using trainable neural networks and perform learning via maximum likelihood estimation (MLE) [14, 26]. However, it can be unstable and expensive due to the need for an exact ODE solution and its differentiation per iteration during training.

Flow matching. Later works attempt to learn CNFs without per-iteration ODE solving [4, 68], evolving into the recent flow-matching (FM) framework [2, 32, 52, 54, 58, 77]. FM moves away from MLE and instead matches a parameterized vector field u_t^θ to a prescribed velocity field u_t that induces a flow to transform $p_0 = p$ into $p_1 = q$. Specifically, for a prescribed probability path $\{p_t\}_{t=0}^1$ and its inducing vector field u_t, FM learns via

$$\min_\theta \ \mathbb{E}_{t,X_t}\|u_t^\theta(X_t) - u_t(X_t)\|^2. \tag{2.2}$$

However, u_t is rarely tractable in practice. Hence, FM considers the *conditional* probability paths $\{p_{t|1}\}_{t=0}^1$ where $p_{t|1} \doteq p(X_t|X_1)$ and the associated *conditional* velocity fields $u_t(X_t|X_1)$, leading to the *conditional* FM loss:

$$\min_\theta \mathbb{E}_{t\sim[0,1],X_0\sim p,X_1\sim q}\|u_t^\theta(X_t) - u_t(X_t|X_1)\|^2, \tag{2.3}$$

whose loss is equivalent to that in Eq. (2.2) up to an additive constant. Eq. (2.3) is tractable in practice. A popular design of $\{p_t\}_{t=0}^1$ is induced by the *linear path*: $X_t = (1-t)X_0 + tX_1$ for $t \in [0,1]$, leading to the *linear conditional flow* [53]. While X_0 and X_1 are usually taken independently, in practice, they can be sampled from any *data coupling* π, i.e., joint distribution $\pi(x_0, x_1)$ with marginals p and q. Under the linear conditional flow, the conditional velocity field is simply given by $u_t(x|X_1) = X_1 - X_0$, so the conditional FM loss in Eq. (2.3) reduces to:

$$\mathcal{L}_{\mathrm{CFM}}^{OT}(\theta) \doteq \mathbb{E}_{t,(X_0,X_1)\sim\pi} \left\| u_t^\theta(X_t) - (X_1 - X_0) \right\|^2. \quad (2.4)$$

Diffusion models (DMs) are an important class of flow-based models with close connections to FM [46, 53]; **We default to FM models in our subsequent exposition**.

Latent flow-based models. In practice, both DMs and FM models are often trained in a latent space with (much) lower dimensions than that of the original space [55, 67]. For this purpose, a pretrained encoder-decoder pair $(\mathcal{E}, \mathcal{D})$—called the visual tokenizer, is responsible for mapping back and forth between the original and the latent spaces. To be precise, given a training set $\{x^i\}_{i=1}^n$, training happens on latent training samples $\{z_i \doteq \mathcal{E}(x^i)\}_{i=1}^n$. During inference, samples are first drawn in the latent space and then mapped back to the original space via $\mathcal{D}$. **Henceforth, we use Z_0 and Z_1 to denote the source and target random variables in the latent space**.

2.2. Training Efficiency of FM Models

FM training efficiency is dictated by the complexity of the target distribution, the source-target distribution distance, and the flow design, among other factors.

Complexity of the target distribution. For latent FM, given a training set $\{x^i\}_{i=1}^n$, the complexity of the latent distribution is solely determined by the tokenizer [7, 20, 81]. In this regard, VAE-based tokenizers (e.g., LDM-VAE [67], SD-VAE [20]) align the latent distribution with a full-dimensional isotropic Gaussian. Since the sample complexity of density estimation scales exponentially with the *effective dimension* at best [9, 57, 60], VAE-based latent FM models tend to suffer from the curse of dimensionality when dealing with full-dimensional latent distributions.

To break the curse, a plausible remedy is to encourage the latent distribution to reveal manifold structures. This is inspired by the celebrated *manifold hypothesis* [3, 5, 19, 21, 56, 65], and in particular, the lifted *Union of Manifolds Hypothesis* (UMH), which states that high-dimensional image data do not lie on a single manifold but on a union of manifolds with *varying* intrinsic dimensions [8, 65]. In line with this remedy, recent visual tokenizers apply vision foundation models to improve latent FM performance through strategies including *alignment* [6, 10, 85], *adaptation* [24, 28], *distillation* [70] and *replacement* [89]. Despite their distinct motivations, we believe that the performance gain comes from: (1) revealing the intrinsic union-of-manifold structures in the latent distribution, validated by superior linear probing accuracy [10, 24, 70, 89]; and (2) the low complexity of the resulting latent distribution, evidenced by the lower fitting loss of a simple density estimation model in the latent distribution as a direct computational proxy for target distribution complexity [11]. Thus, to reduce the complexity of the target distribution, one needs to choose a visual tokenizer empowered by visual foundation models that reveal the union-of-manifold structures.

Distance between source and target. Once the target distribution is fixed, the next critical factor is the source distribution p. For this, p as the isotropic Gaussian has been a standard choice. But, intuitively, to minimize the transport efforts of transforming the source into the target, we want to make the source-target distance small [53]. In this vein, recent works propose using target-approximating source distributions to replace the isotropic Gaussian source [27, 37, 41, 61, 75] in specific domains and observe substantial performance gains. For example, Mirror FM [27] and heavy-tailed DM [61] show that when the target is heavy-tailed, working with a heavy-tailed source is critical to avoiding the divergent velocity field and high transport cost caused by the light-tailed Gaussian source. So, to work with the multimodal, union-of-manifold target distribution induced by foundation-model-based tokenizers, one might want to choose a tractable multimodal distribution as the source.

Flow Design. The complexity of the flow that bridges the source and target distributions also contributes significantly to the learning difficulty, primarily through the choice of path (e.g., affine path, Gaussian path) and data coupling [53]. The *linear* FM in Sec. 2.1 is the most popular among the paths [20, 76, 81] due to its simplicity and mathematical optimality in terms of transport cost [52, 54]. Regarding data coupling, BatchOT [64] constructs non-trivial couplings by solving a mini-batch optimal transport (OT) map between source and target samples, and [48] learns good data coupling alongside the FM training process. When the source is close to the target, we intuitively only need to move probability mass locally. In this case, linear paths and local data coupling are probably sufficient.

2.3. Similar but Different Works

There are two similar but different ideas compared to ours, as detailed in Tab. 1. MixSGM [36] operates in the pixel space and thus is limited to **low-dimensional** EMNIST and CIFAR-10 data. In contrast, our method scales to **high-dimensional** natural images by leveraging vision foundation models to reveal the union-of-manifold structures, in-

Table 1. The comparison of our method with similar methods. †: Covariance refers to explicit covariance estimation instead of assuming for isotropic covariance.

Method	CondPrior [35]	MixSGM [36]	Ours
Latent FM Model	✓	✗	✓
Multimodal Tokenizer	✗	✗	✓
GMM with Covariance†	✗	✗	✓
Local Data Coupling	✗	✓	✓

ducing a low-complexity target distribution for FM. CondPrior [35] applies a VAE-based tokenizer to perform latent FM for ImageNet-1K, and forms the mixture components of their GMM source based on annotated classes. However, the poor ImageNet-1K linear probing results [70, 89] suggest that the VAE-based tokenizer alone fails to cluster ImageNet-1K well. So, such a class-conditioned GMM source distribution might not reflect the true structure of the latent distribution, not to mention that annotated classes are not always available for image datasets.

Remark: The work Gaussian Mixture Flow Matching (GMFlow) [12] shares only a naming similarity to our work. Their motivation lies in using GMM to model the conditional velocity field—intrinsically a distribution. Moreover, we refer the reader to remotely related work in Sec. 9.

3. Method

The design of our algorithm—dubbed `MM-FM`—closely follows the insights in Sec. 2.2 to reduce the learning difficulty: (1) employing a visual tokenizer that reveals the union-of-manifold structure, resulting in a multimodal target distribution; (2) estimating a target-approximating source distribution via GMM; and (3) designing a mode-dependent data coupling for the probability mass to transport locally.

3.1. Our Algorithm

Foundation model as visual tokenizer to reduce the complexity of the target distribution. Because visual foundation models (e.g., DINOv2 [17], SigLIP2 [79]) are known for their superior performance in linear probing, they are good candidates as encoders to reveal the union-of-manifold structures in the latent space. Although many recent works on latent FM leverage foundation models to regularize encoders [6, 10, 24, 70, 85], the linear probing accuracy is not as good as that of using foundation models as encoders directly [28, 89]. Based on empirical linear probing performance and t-SNE visualization (see Sec. 11.1), we adopt the DINOv2-B foundation model as our encoder (see Fig. 1). To train a corresponding decoder, we follow the training recipe in [89] and consider the combined loss

$$\mathcal{L} = \mathrm{L1}(\hat{x}, x) + \omega_L \,\mathrm{LPIPS}(\hat{x}, x) + \omega_G \lambda \,\mathrm{GAN}(\hat{x}, x), \quad (3.1)$$

i.e., a combination of reconstruction (L1), fidelity (LPIPS [88]), and adversarial losses [25], where $z = \mathcal{E}(x)$, $\hat{x} = \mathcal{D}(z)$. The decoder is not required for training `MM-FM` but is necessary for image synthesis (inference).

Narrowing source-target gap with GMM. The union-of-manifold structure implies a multimodal distribution. To reduce the distance between the source distribution and the multimodal target distribution, we fit a GMM to the target distribution and use it as our source distribution. This GMM source distribution can be viewed as a warm-start estimation of the target and FM training as refinement. Since GMM estimation struggles due to the high dimensionality, we design a subroutine to produce an operational GMM (see Sec. 3.2).

Mode-dependent data coupling. The GMM fitting from the last step naturally assigns target samples to their corresponding Gaussian modes. Moving beyond independent coupling, we leverage these assignments to design a mode-dependent coupling termed *mode coupling*. Specifically, given $z_1 \sim q$, we softly assign z_1 to the Gaussian modes of p via their posterior responsibilities, and define the mode coupling by sampling z_0 from the resulting mixture:

$$\pi_{\text{mode}}(z_0 \mid z_1) = \sum_{k=1}^{m} w_k(z_1)\, \mathcal{N}(z_0;\, \mu_k,\, \Sigma_k), \quad (3.2)$$

$$w_k(z_1) = \frac{c_k\, \mathcal{N}(z_1;\, \mu_k,\, \Sigma_k)}{\sum_{i=1}^{m} c_i\, \mathcal{N}(z_1;\, \mu_i,\, \Sigma_i)}. \quad (3.3)$$

The complete algorithms are in Algorithms 1 and 2.

Algorithm 1 Training our `MM-FM`

Input: an image dataset $\{x_1^{(j)}\}$
Output: a velocity field neural network u_t^θ

Stage 0: Vision Foundation Model Encoding
$\{z_1^{(j)}\} \leftarrow \mathcal{E}\{x_1^{(j)}\}$ ▷ *Encode images to multimodal latents*

Stage 1: Gaussian Mixture Model (GMM) Fitting
$m \leftarrow$ Bayesian GMM$(\{z_1^{(j)}\})$ ▷ *Infer number of modes*
Fit GMM $p = \sum_{i=1}^{m} c_i \mathcal{N}(\mu_i, \Sigma_i)$ to $\{z_1^{(j)}\}$

Stage 2: Flow Matching Training
Initialize parameters θ for u_t^θ
for each training iteration **do**
  Sample $(Z_0, Z_1) \sim \pi_{\text{mode}}(z_0|z_1)$ via mode coupling:
    Sample $Z_1 \sim q$ ▷ *Sample data*
    Sample mode k from posterior $p(\cdot \mid Z_1; \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
    Sample $Z_0 \sim \mathcal{N}(\mu_k, \Sigma_k)$ ▷ *Sample noise*
  Sample $t \sim \mathcal{U}(0,1)$ and set $Z_t = (1-t)Z_0 + tZ_1$
  Descend gradient on loss $\mathcal{L}_{\text{CFM}}^{OT}(\theta)$:
    $\nabla_\theta \| u_t^\theta(Z_t) - (Z_1 - Z_0) \|^2$

Relation to CondPrior [35] and MixSGM [36]. They do not utilize a foundation-model-based encoder to reveal the multimodal target distribution. As discussed in Sec. 3.1, the encoding stage of Algorithm 1 is crucial for revealing the

Algorithm 2 Sampling from MM−FM

Input: trained u_t^θ, fitted GMM $\{c_i, \mu_i, \Sigma_i\}_{i=1}^m$, ODE steps N
Sample $k \sim \mathrm{Cat}(c_1, \ldots, c_m)$, then $z_0 \sim \mathcal{N}(\mu_k, \Sigma_k)$
for $n = 0, \ldots, N-1$ **do** ▷ *Euler over* $t \in [0, 1]$
  $z_{(n+1)/N} \leftarrow z_{n/N} + \frac{1}{N} \cdot u^\theta_{n/N}(z_{n/N})$
return $\hat{x}_1 \leftarrow \mathcal{D}(z_1)$ ▷ *decoder for foundation model* Enc

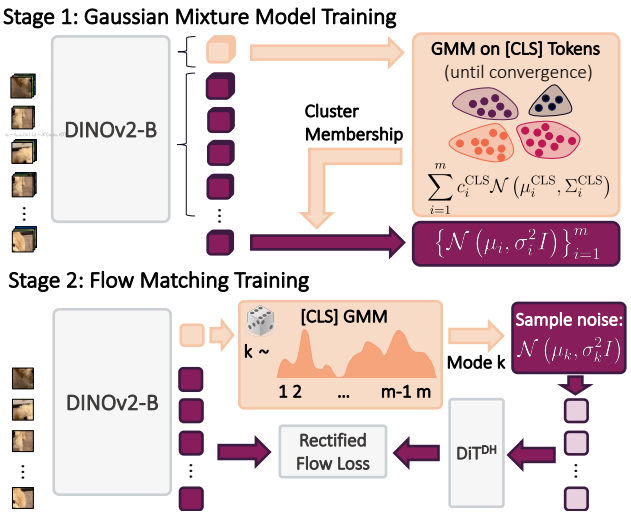

Figure 2. We leverage the [CLS]-token space ($\mathbb{R}^{768}$) and the one-to-one correspondence between patch and [CLS] tokens to produce an operational GMM in the patch-token space.

union-of-manifold structures and is also a necessary condition for the estimation of the GMM source distribution.

3.2. GMM Implementation in High Dimensions

Although our choice of the DINOv2-B model as the encoder induces a low-complexity multimodal distribution, the latent dimension is very high, defying effective GMM estimation. Specifically, DINOv2-B [17, 89] encodes each image into 768 patch tokens, each 16×16 in size, plus a [CLS] token of size 768. To break the curse of dimensionality, we leverage this (*relatively*)-low-dimensional [CLS] token, a global descriptor of its corresponding patch tokens, to derive an operational GMM estimation procedure for the high-dimensional latent distribution in $\mathbb{R}^{768\times16\times16}$.

As shown in Fig. 2, we first run a full-fledged GMM estimation in the [CLS] token space to determine cluster memberships; then, we estimate the Gaussian mean and variance of the patch tokens for each cluster separately in the patch-token space. During FM training, we pair an image latent (i.e., patch tokens) with noise drawn from a specific mode by (1) selecting a mode via sampling from the soft assignment probabilities for the observed [CLS] token in the [CLS]-token space, and then (2) drawing noise with the mean and variance of that particular Gaussian mode in the patch-token space. During FM generation, we sample mode indices from the [CLS] GMM and draw noise samples from the corresponding Gaussian modes in patch-token space.

3.3. Toy Experiment

Table 2. Results of a toy experiment on $\mathbb{R}^{10}$. **Len** is the average trajectory length and W_2^2 is the square of 2-Wasserstein distance. Using GMM source alone is insufficient, but co-designed with mode coupling MM−FM obtains superior performance.

Source	Gaussian		GMM			
Coupling	Indep.		Indep.		MC	
Data↓	**Len**↓	W_2^2↓	**Len**↓	W_2^2↓	**Len**↓	W_2^2↓
Compact	1.88	1.252	1.82	1.251	**1.67**	**1.248**
Normal	1.97	1.366	1.92	1.368	**1.72**	**1.364**
Spread	2.35	1.582	2.30	1.631	**2.00**	**1.556**

To verify that our MM−FM reduces the learning difficulty, we run a toy experiment in $\mathbb{R}^{10}$, where the length of the sampling trajectory and the 2-Wasserstein distance can be computed efficiently. In Tab. 2, the shorter average sampling trajectory (**Len**) implies that MM−FM only requires local transport of probability mass, and the lower 2-Wasserstein distance (W_2^2) indicates reduced learning difficulty. We also note that (1) MM−FM brings in straighter sampling trajectories that also improve inference efficiency; and (2) MM−FM is compatible with BatchOT [64] if computing allows for further improvements in trajectory length, trajectory straightness, and 2-Wasserstein distance. We refer the reader to Sec. 7 for complete details of this experiment.

3.4. Theoretical Analysis

We take a first cut to mathematically justify the improvement yielded by our method MM−FM. To make the analysis tractable, we introduce some simplifying assumptions. Specifically, we assume that the target density q is a uniform mixture of affine translations of a fixed density $\tilde{q}$:

$$q(z) = \frac{1}{m}\sum_{k=1}^m \tilde{q}(\Sigma_k^{-1/2}(z-\mu_k)) = \frac{1}{m}\sum_{k=1}^m q_k(z),$$

where $\tilde{q}$ is a unimodal probability density on $\mathbb{R}^d$, $\{\mu_k\}_{k=1}^m \subset \mathbb{R}^d$ are mean vectors, $\{\Sigma_k\}_{k=1}^m \subset \mathbb{R}^{d\times d}$ are covariance matrices, and $q_k(z) \doteq \tilde{q}(\Sigma_k^{-1/2}(z-\mu_k))$. To standardize the problem, we assume $\mathbb{E}_{\tilde{q}}[z] = 0$ and $\mathbb{E}_{\tilde{q}}[zz^T] = I_d$. We refer to producing samples from $\tilde{q}$ as the **unimodal generation problem** (UGP), and those from q as the **multimodal generation problem** (MGP).

Our goal is to show that an informed source and coupling choice for the MGP can make ease both learning and generation processes. First, consider another unimodal distribution $\tilde{p}(z)$ satisfying $\mathbb{E}_{\tilde{p}}[z] = 0$ and $\mathbb{E}_{\tilde{p}}[zz^T] = I_d$, called the **unimodal source**. Then, we define

$$p(z) = \frac{1}{m}\sum_{k=1}^m \tilde{p}(\Sigma_k^{-1/2}(z-\mu_k)) = \frac{1}{m}\sum_{k=1}^m p_k(z)$$

to be the **multimodal source**, with the same parameters $\{\mu_k, \Sigma_k\}_{k=1}^m$ as those of the target distribution q. Thus, p and q have the same modes and a similar "shape" around each mode. This makes the distance between p and q smaller than that between $\tilde{p}$ and q. Moreover, we introduce three quantities that measure the complexity of the flow:

1. The *straightness* of flow, defined by

$$\mathbb{E}\left[\int_0^1 \|u_t(Z_t)\|^2 dt - \left\|\int_0^1 u_t(Z_t)dt\right\|^2\right].$$

   Intuitively, the two terms measure the total transport work and effective transport work done, respectively. The smaller the difference is, the straighter the flow is. Thus, a small straightness value is favored in our setting, as it suggests that the flows are closer to linear and can be simulated with fewer integration steps.

2. The *total length* of flow, defined by

$$\text{Len} = \mathbb{E}\left[\int_0^1 \|u_t(Z_t)\| dt\right].$$

   This quantity captures the average distance that the flow takes to bridge a particle X_0 from the source distribution to a particle X_1 from the target distribution. In general, a small total length is favorable.

3. The *Lipschitz constant of the velocity field*, given by

$$\text{Lip}(u_t) = \sup_z \|\nabla u_t(z)\|_{\text{op}}.$$

   The Jacobian of the velocity field, $\nabla u_t(z)$, measures how fast the vector field changes with respect to position; thus, the Lipschitz constant is a direct measure of the curvature of the flow. The Lipschitz constant also captures the difficulty of learning the velocity field with neural network models, as a large Lipschitz constant indicates a highly oscillatory vector field, which is less amenable to neural network approximation. In particular, empirically, neural networks may struggle to learn highly oscillatory functions due to spectral biases [66]. Thus, a smaller Lipschitz constant is favorable.

To simplify the analysis, we assume the mode coupling π_{mode} is defined using the "hard" version of Eq. (3.3)[2], i.e., $w_k(z_1) = \mathbb{1}\left[k = \arg\max_{j\in[m]} \mathcal{N}(z_1; \mu_j, \Sigma_j)\right]$, equivalent to the mode assignment function $k : \mathbb{R}^d \to [m]$ by:

$$k(z) = \arg\min_{k\in[m]} \|z - \mu_k\|_{\Sigma_k^{-1/2}}.$$

In addition, we place some technical assumptions on the distributions $\tilde{p}$ and $\tilde{q}$, $\{\mu_k\}_{k=1}^m$, and $\{\Sigma_k\}_{k=1}^m$.

Assumption 1. *1. For both $\tilde{p}$ and $\tilde{q}$, the support Ω is a bounded, convex subset of $\mathbb{R}^d$, and its affine images $\{\Omega_k\}_{k\in[m]}$ are mutually disjoint. In addition, for every pair of distinct indices $j, k \in [m]$, and every $z \in \Omega_k$,*

$$\|z - \mu_k\|_{\Sigma_k^{-1/2}} < \textit{dist}_{\Sigma_j^{-1/2}}(z, \Omega_j),$$

where $\textit{dist}_{\Sigma_j^{-1/2}}(z, \Omega_j) = \inf_{y\in\Omega_j} \|z - y\|_{\Sigma_j^{-1/2}}$.

2. The considered velocity field $\tilde{u}_t$ bridging the unimodal distributions $\tilde{p}$ and $\tilde{q}$ is Lipschitz continuous for every $t \in (0, 1)$, and its Jacobian is defined globally in Ω.

The first item of Assumption 1 essentially states that the modes $\{\mu_k\}_{k=1}^m$ are sufficiently well-separated. The second item of Assumption 1 is a technical assumption on the $\tilde{p}$ and $\tilde{q}$, which is satisfied when both $\tilde{p}$ and $\tilde{q}$ are smooth and their ratio $\frac{\tilde{p}}{\tilde{q}}(x)$ is always nonzero; in other words, wherever $\tilde{p}(x)$ places positive mass, $\tilde{q}(x)$ also places positive mass. Our main result is stated below:

Theorem 3.1. *Let Assumption 1 hold. If we implement FM on the multimodal distribution q with the multimodal source p and the mode coupling π_{mode}, the following holds.*

1. The straightness is bounded by that of the UGP:

$$\textit{Straightness}(p, q; \pi_{mode}) \le C \cdot \textit{Straightness}(\tilde{p}, \tilde{q}; \pi_{ind}),$$

where $C := \left(\frac{1}{m}\sum_{k=1}^m \|\Sigma_k\|_{op}\right) \le 1$.

2. The total length is bounded by that of the UGP:

$$\textit{Len}(p, q; \pi_{mode}) \le \left(\frac{1}{m}\sum_{k=1}^m \|\Sigma_k\|_{op}^{1/2}\right) \cdot \textit{Len}(\tilde{p}, \tilde{q}; \pi_{ind}).$$

3. If u_t denotes the velocity field bridging p and q under the mode coupling, and $\tilde{u}_t$ denotes the velocity field bridging $\tilde{p}$ and $\tilde{q}$ under the independent coupling, we have

$$\textit{Lip}(u_t) \le \textit{Lip}(\tilde{u}_t),\ \forall t \in (0, 1).$$

See Sec. 6 for the proof details. Theorem 3.1 states that when using the multimodal source and the mode coupling, the straightness, total length, and the Lipschitz constant of the velocity field of the UGP *never exceed those of the UGP*. We remark that the constants $\frac{1}{m}\sum_{k=1}^m \|\Sigma_k\|_{\text{op}}$ and $\frac{1}{m}\sum_{k=1}^m \|\Sigma_k\|_{\text{op}}^{1/2}$ may be pessimistic, as they arise from worst-case estimates of the form $\|\Sigma u_t(z)\| \le \|\Sigma\|_{\text{op}}\|u_t(z)\|$. Thus, equality holds in items 1 and 2 of Theorem 3.1 only when these bounds are sharp. When the covariances $\{\Sigma_k\}_{k=1}^m$ have small eigenvalues in certain directions, the constant may be substantially improved.

To further demonstrate the importance of a multimodal source for a multimodal target, we include a complementary result describing the limitations of a unimodal source.

Theorem 3.2. *Let $\tilde{p}$ denote the unimodal source and q denote the multimodal target. Let u_t denote the velocity field bridging $\tilde{p}$ and q under the independent coupling and let L_t*

[2] In high-dimensional settings where the concentration of measure phenomenon holds, the soft and hard mode assignments behave similarly.

Table 3. **System-level comparison of latent flow-based unconditional image generation on ImageNet 256×256**. The class-conditioned methods are only for reference, not for comparisons. †: Results are taken from RAE manuscript [89]. Our DiT^{DH}-XL use 50 ODE steps, while SVG-XL uses 25 steps and the remaining 250 steps.

Method	Tokenizer	Training Epoches	# params	Prior	Condition	Generation w/o guidance				Generation w/ guidance			
						FID	IS	Pre.	Rec.	FID	IS	Pre.	Rec.
LDM-8 [67]	LDM	150	395M	Gaussian	-	39.13	-	-	-	-	-	-	-
DiT-XL [51]	SD-VAE	400	675M	Gaussian	-	27.32	35.90	-	-	-	-	-	-
DiT-XL [62]	SD-VAE	80	675M	Gaussian	Class	19.50	-	-	-	-	-	-	-
SiT-XL [55]	SD-VAE	80	675M	Gaussian	Class	17.20	-	-	-	-	-	-	-
SiT-XL [49]	E2E-VAE	80	675M	Gaussian	Class	3.46	159.8	0.77	0.63	1.67	266.3	0.80	0.63
LightningDiT-XL [85]	VA-VAE	64	675M	Gaussian	Class	5.14	130.2	0.76	0.62	2.11	252.3	0.81	0.58
LightningDiT-XL [11]	MAETok	64	675M	Gaussian	Class	5.36	-	-	-	3.24	-	-	-
LightningDiT-XL [10]	AlignTok	64	675M	Gaussian	Class	3.71	148.9	0.77	0.62	1.90	260.9	0.81	0.61
LightningDiT-XL [10]	VFM-VAE	80	675M	Gaussian	Class	3.41	160.4	-	-	-	-	-	-
SVG-XL [70]	SVGTok	80	675M	Gaussian	Class	6.57	137.9	-	-	3.54	207.6	-	-
DiT-XL + RCG [51]	SD-VAE	400	738M	Gaussian	MoCov3	4.89	143.20	-	-	-	-	-	-
DiT-XL + DLC$_{512}$ [47]	SD-VAE	80	825M	Gaussian	DLC	5.75	-	-	-	-	-	-	-
DiT^{DH}-XL	**DINOv2-B**	20	839M	Gaussian	-	16.51	61.36	0.69	0.62	12.45	71.36	0.72	0.63
		80	839M	Gaussian	-	9.33	90.62	0.71	0.65	5.82	110.19	0.73	0.66
		200	839M	Gaussian	-	-	-	-	-	4.96†	123.12†	-	-
DiT^{DH}-XL + MM-FM	**DINOv2-B**	20	839M	**GMM**	-	5.11	188.37	0.81	0.49	4.39	182.59	0.82	0.52
		80	839M	**GMM**	-	**3.82**	192.28	0.81	0.54	**3.23**	183.62	0.81	0.57
		20	839M	**GMM**	**Mode**	4.74	194.76	0.82	0.48	4.20	184.18	0.83	0.51
		80	839M	**GMM**	**Mode**	**3.18**	211.23	0.83	0.53	**2.74**	197.27	0.83	0.56

denote the Lipschitz constant of u_t. Then for any constant $M > 0$, there exists a choice of $\{\mu_k, \Sigma_k\}_{k=1}^m$ such that the averaged Lipschitz constant satisfies

$$\int_0^1 L_t \geq M.$$

See Sec. 6 for the proof details. Theorem 3.2 states that under the unimodal source and the independent coupling, the velocity field's Lipschitz constant can grow arbitrarily large due to scale imbalance between prior and target.

To summarize, our theory explains the importance of the co-design of the coupling and source distribution when learning multimodal distributions with FM. Theorem 3.1 shows that such co-design eases the learning process by making the velocity field less curved and also expedites the generation process by promoting straighter flows. Complementing Theorem 3.1, Theorem 3.2 shows that with a unimodal source, the velocity field can become *arbitrarily curved*. Our key theoretical insight is that the multimodal source, coupled with mode coupling, ensures that components with the same covariance structure are matched and flows stay within a single component of the support.

4. Experiments

Setup. We conduct experiments on ImageNet-256 [44], containing 1.28M images. Following the standard protocol established by ADM [18], we pre-process all images by center-cropping to maintain aspect ratio and resizing to 256×256 resolution. During training, we apply horizontal flipping as the sole data augmentation. We employ DiT^{DH}-XL as the seminal diffusion transformer backbone operating directly in raw structured DINOv2-B latents, and we also strictly follow its learning settings [89]. Unless specified, we default to use GMM with 8192 modes, diagonal covariance types, and soft assignment. The time spent training the GMM model (once) is negligible compared to training the flow matching model. Our GPU implementation of using the trained GMM for mode coupling almost adds no overhead to the training speed compared to independent coupling with isotropic Gaussian. The detailed runtime analysis and training parameters are in Sec. 10.

Conditioning and sampling strategies. Unlike prior works on class-conditioned generation [70, 85, 89], we primarily focus on the scenario without the luxury of supervised labels to make better exploit foundation models and data structures. We also set our training budget to be 80 epochs for resource-constrained settings where foundation models are increasingly practical. We include a variant of `MM-FM` that takes the mode index as an additional conditioning input, which helps guide samples at mode boundaries. In Sec. 8, we show that *the minimizer of the OT objective Eq. (2.4) approximately coincides with the minimizer of the mode-conditional variant*. Although classifier-free guidance does not apply to the unconditional setting, AutoGuidance remains applicable [38], where the detailed guidance parameters are in Sec. 10. Following [89], our baseline for apple-to-apple comparison, we use an ODE sampler with 50 Euler steps unless specified. We also include RCG [51] and DLC [47] as current SOTAs for unconditional ImageNet-256 generation.

Evaluation protocol. Consistent with our baselines, all models are evaluated on 50,000 generated samples. Our primary metric is FID, with Inception Score (IS), precision, and recall [33, 45, 69]. Reference statistics are ADM's pre-computed values [18] over the full ImageNet dataset.

4.1. State-of-the-art Unconditional DiT

As shown in Tab. 1, our MM−FM (with autoguidance) establishes new SOTA unconditional generation (FID=2.74) on ImageNet-256 within 80 epochs. Under such a resource-constrained setup, our FID scores (without guidance) of both variants surpass RCG with full training and outperform DLC by 45%, without training an additional generative model to provide representation conditions [47, 51]. Notably, without guidance, our MM−FM even matches and surpasses many class-conditioned methods, making it suitable for large-scale unsupervised datasets. Even without explicit mode conditioning, MM−FM still obtains a competitive score of FID 3.82 (2.44× better than the classic FM recipe).

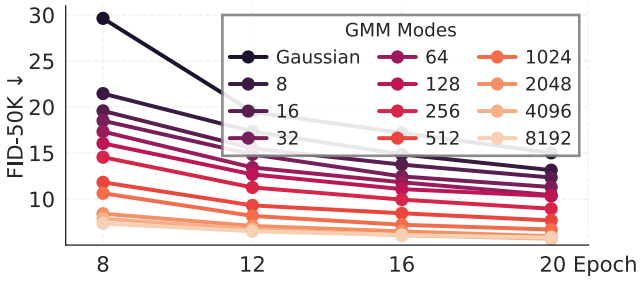

Figure 3. The ablative study of mode numbers for MM−FM without mode conditioning on unconditional image generation. The GMMs are using diagonal covariances and soft assignment.

4.2. MM-FM Exploits the Structured Latents

As shown in Fig. 3, MM−FM strictly improves the training efficiency (faster convergence) relative to Gaussian source with independent coupling, and we observe a monotonic improvement with increasing number of modes toward the degree of reasonable GMM estimation with sufficient data points per mode. At Epoch 8, we clearly observe the "warm-up" effects brought by MM−FM. Notably, the benefits persist even when the number of modes exceeds the number of class labels in ImageNet, demonstrating that class labels do not represent or fully exploit the true underlying structure of the latent space. This suggests that the data manifold contains richer intrinsic structure beyond the categorization provided by supervised labels, which MM−FM can effectively capture. Let alone that class labels are not always available. We defer the ablation study of covariance and assignment types in Sec. 11 as we empirically find that they only marginally affect the generation performance. As a side note, the GMM fitting does not need to be perfect because any reasonable multimodal structures learned are relatively better than a target-blind isotropic Gaussian.

4.3. Less Crossing and Straighter Trajectories

In the few-step generation regime, our MM−FM only employs 5 ODE steps relative to classic FM requiring 25 ODE steps to achieve the same FID score. Because a straight trajectory needs fewer steps than a curve one to achieve a target discretization error, MM−FM reveals straighter sampling trajectories, greatly improving the inference efficiency.

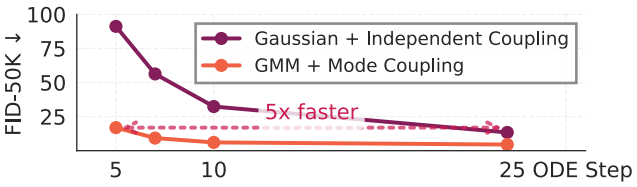

Figure 4. MM−FM demonstrates outstanding inference efficiency. The GMM $m = 8192$ uses diagonal covariance and soft assignment and MM−FM without mode conditioning.

4.4. Enable Data-limited Generative Modeling

Beyond training and inference efficiency, MM−FM demonstrates outstanding data efficiency, outperforming, outperforming classic FM by a greater margin than in the full-data regime (see Tab. 4). We attribute this to MM−FM lowering the learning complexity, thus reducing demand for training data. In particular, we train both MM−FM and classic FM toward convergence, so the significant gap in the FID scores really demonstrates the reduction of the learning problem complexity by allowing probability mass to move locally.

Table 4. Training DiT^{DH}-S on stratified 10% of ImageNet256 data with the baseline (Gaussian source with independent coupling) and MM−FM with 8192 modes, where the GMM is also estimated from the 10% ImageNet data.

	400 Epochs		800 Epochs	
FID-50K ↓	Baseline	MM−FM	Baseline	MM−FM
	24.65	**8.04**	24.33	**7.48**

5. Conclusion

In this paper, we propose MM-FM, a co-design of multimodal source distribution and mode-dependent coupling that exploits structured foundation model latents, achieving SOTA unconditional generation with 30× faster convergence through straighter, shorter trajectories and lower learning complexity. We believe that MM−FM represents a significant leap into the co-evolution of foundation models and generative models: as foundation models emerge across domains [30, 71], flow-based generative modeling can be made efficient by fully exploiting these semantically-rich structured representations and building localized flows.

Acknowledgments

Luo G. was partially supported by a UMN DSI-MnDRIVE PhD Graduate Assistantship. Sun. J was partially supported by a UMN DSI Faculty Fellowship. This work was partially supported by NIH R01CA287413. Lu Y. was supported by NSF CAREER Award DMS-2442463. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. The co-authorship order was determined by a coin flip. Luo G. and Cole F. contributed equally and reserve the right to list their name first in their resumes.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1

[2] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[3] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *ArXiv*, abs/1701.07875, 2017. 2, 3

[4] Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Maximilian Nickel, Aditya Grover, Ricky T. Q. Chen, and Yaron Lipman. Matching normalizing flows and probability paths on manifolds. In *International Conference on Machine Learning*, 2022. 2

[5] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2012. 3

[6] Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models, 2025. 2, 3, 4, 24, 27

[7] Black Forest Labs. Flux, 2024. 1, 2, 3, 24, 27

[8] Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data, 2023. 2, 3

[9] Theophilos Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966. 3

[10] Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models, 2025. 2, 3, 4, 7, 24

[11] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. 2, 3, 7, 24

[12] Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models. In *Forty-second International Conference on Machine Learning*, 2025. 4, 24

[13] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 24

[14] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 6572–6583, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[15] Yifan Chen, Eric Vanden-Eijnden, and Jiawei Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025. 2

[16] Max Daniels. On the contractivity of stochastic interpolation flow. *arXiv preprint arXiv:2504.10653*, 2025. 16

[17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 5, 24

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 7, 8, 26

[19] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016. 3

[20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 2, 3, 24

[21] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 3

[22] Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024. 2

[23] Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of machine learning research*, 26(43):1–54, 2025. 2

[24] Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation, 2025. 2, 3, 4, 24

[25] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[26] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019. 2

[27] Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains, 2025. 3

[28] Ming Gui, Johannes Schusterbauer, Timy Phan, Felix Krause, Joshua M. Susskind, Miguel Ángel Bautista, and Björn Ommer. Adapting self-supervised representations as a latent space for efficient generation. In *The Fourteenth International Conference on Learning Representations*, 2026. 3, 4, 24

[29] Pengsheng Guo and Alex Schwing. Variational rectified flow matching, 2025. 24

[30] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. 8

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 27

[32] Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α -(de)blending: Blending: a minimalist deterministic diffusion model. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[33] 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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 8, 26

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 24

[35] Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models, 2025. 4, 23

[36] Nanshan Jia, Tingyu Zhu, Haoyu Liu, and Zeyu Zheng. Structured diffusion models with mixture of gaussians as prior distribution, 2024. 3, 4, 23

[37] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 3

[38] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7, 26

[39] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proc. CVPR*, 2024. 24

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 24

[41] Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[42] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025. 24

[43] Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 24

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc. 7

[45] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. 8, 26

[46] Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models, 2025. 1, 3

[47] Samuel Lavoie, Michael Noukhovitch, and Aaron Courville. Compositional discrete latent code for high fidelity, productive diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 7, 8

[48] Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ode-based generative models. In *International Conference on Machine Learning*, pages 18957–18973. PMLR, 2023. 3

[49] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18262–18272, 2025. 7, 24

[50] Fanfei Li, Thomas Klein, Wieland Brendel, Robert Geirhos, and Roland S. Zimmermann. LAION-c: An out-of-distribution benchmark for web-scale vision models. In *Forty-second International Conference on Machine Learning*, 2025. 27

[51] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7, 8, 25

[52] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[53] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. 1, 3, 24

[54] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 24

[55] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow diffusion-based generative models with scalable interpolant transformers. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, page 23–40, Berlin, Heidelberg, 2024. Springer-Verlag. 3, 7, 24

[56] James R Munkres. *Analysis on manifolds*. CRC Press, 2018. 3

[57] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2010. 3

[58] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: learning stochastic dynamics from samples. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2

[59] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 27

[60] Arkadas Ozakin and Alexander Gray. Submanifold density estimation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009. 3

[61] Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[62] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 7, 24, 25

[63] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 26, 27

[64] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3, 5, 22

[65] Phillip E. Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *ArXiv*, abs/2104.08894, 2021. 2, 3

[66] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019. 6

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3, 7

[68] Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In *Advances in Neural Information Processing Systems*, 2021. 2

[69] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. 8, 26

[70] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder, 2025. 2, 3, 4, 7, 24, 27

[71] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 8, 27

[72] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2026. 24

[73] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 2256–2265. JMLR.org, 2015. 24

[74] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 24

[75] Hannes Stark, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic prior self-conditioned flow matching for multi-ligand docking and binding site design. In *NeurIPS 2023 AI for Science Workshop*, 2023. 3

[76] SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images, 2025. 3

[77] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.*, 2024, 2024. 2, 22

[78] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint*, 2026. 28

[79] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 4, 27, 28

[80] Panos Tsimpos, Zhi Ren, Jakob Zech, and Youssef Marzouk. Optimal scheduling of dynamic transport. *arXiv preprint arXiv:2504.14425*, 2025. 2

[81] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. 1, 2, 3, 24

[82] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer, 2025. 24

[83] Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. REPA works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 24

[84] Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18770–18780, 2025. 24

[85] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 4, 7, 24, 25, 27

[86] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. 24

[87] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. 24

[88] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[89] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025. 2, 3, 4, 5, 7, 24, 25, 26, 27, 28