6. Proofs

We begin with a crucial lemma which defines an equivalent characterization of the mode coupling under Assumption 1.

Lemma 6.1. *Define a sampling procedure as follows:*
1. Sample a random index from $[m]$ such that $\mathbb{P}(k) = c_k$;
2. Conditioned on k, sample $Z_0' \sim p_k$ and $Z_1' \sim q_k$ independently.
Denote by π_* the joint law of (Z_0', Z_1'). Then $\pi_* = \pi_{mode}$.

Proof of Lemma 6.1. It is clear that the marginal distribution of Z_0' is p and the marginal distribution of Z_1' is q, so we need only check that the conditional distribution of $\pi_*(Z_0'|Z_1')$. Note that

$$\begin{aligned}\pi*(Z_0'|Z_1') &= \pi_*(Z_0' \in \Omega_k|Z_1')\pi_*(Z_0'|Z_1',\ Z_0 \in \Omega_k) \\ &= \mathbf{1}(Z_1' \in \Omega_k)p_k(Z_0'),\end{aligned}$$

where we used the definition of π_*. On the other hand,

$$\begin{aligned}\pi_{\text{mode}}(Z_0|Z_1) &= \pi_{\text{mode}}(Z_0 \in \Omega_k|Z_1)\pi_{\text{mode}}(Z_0|Z_1,\ Z_0 \in \Omega_k) \\ &= \mathbf{1}(k \in \text{argmin}_{j\in[m]}\|Z_1 - \mu_k\|_{\Sigma_k^{-1/2}})p_k(Z_0).\end{aligned}$$

Thus, it suffices to show that $x \in \Omega_k \iff k \in \text{argmin}_{j\in[m]}\|z - \mu_j\|_{\Sigma_j^{-1/2}}$, which follows from Assumption 1. Indeed, if $z \in \Omega_k$, then for every $j \neq k$,

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

where the first inequality is due to Assumption 1; hence $k \in \arg\min_{j\in[m]} \|z - \mu_j\|_{\Sigma_j^{-1/2}}$. On the other hand, suppose that $k \in \arg\min_{j\in[m]} \|z - \mu_j\|_{\Sigma_j^{-1/2}}$ but $z \notin \Omega_k$ (say, $z \in \Omega_{k'}$ for some $k' \in [m]$). Then Assumption 1 implies that

$$\|z - \mu_{k'}\|_{\Sigma_{k'}^{-1/2}} \leq \text{dist}_{\Sigma_k^{-1/2}}(z, \Omega_k) \leq \|z - \mu_k\|_{\Sigma_k^{-1/2}},$$

which contradicts the optimality of $\|z - \mu_k\|_{\Sigma_k^{-1/2}}$; hence we must have $z \in \Omega_k$. We conclude the proof. □

Next, we use Lemma 6.1 to derive an expression for the velocity field between p and q under the mode coupling.

Lemma 6.2. *Let u_t denote the velocity field u_t between p and q under π_{mode} and, for $k \in [m]$, let $u_t^{(k)}$ denote the velocity field between p_k and q_k under the independent coupling. Then, under Assumption 1, $u_t(\cdot)$ is a map $(0,1) \times \bigcup_{k=1}^m \Omega_k \to \mathbb{R}^d$ such that*

$$u_t(z) = u_t^{(k)}(z),\ \forall z \in \Omega_k.$$

Proof. By Lemma 6.1, we can sample (Z_0, Z_1) from π_{mode} by first sampling an index k and then sampling from p_k and q_k independently. It follows that

$$\begin{aligned}u_t(z) &= \mathbb{E}_{(Z_0,Z_1)\sim\pi_{\text{mode}}}[Z_1 - Z_0|Z_t = z] \\ &= \mathbb{E}_K\mathbb{E}_{(Z_0,Z_1|K)}[Z_1 - Z_0|Z_t = z, K] \\ &= \sum_{k=1}^m \pi_{\text{mode}}(K = k|Z_t = z)\mathbb{E}_{(Z_0,Z_1)\sim p_k\otimes q_k}[Z_1 - Z_0|Z_t = z] \\ &= \sum_{k=1}^m \pi_{\text{mode}}(K = k|Z_t = z)u_t^{(k)}(z).\end{aligned}$$

By Bayes' rule, we have

$$\pi_{\text{mode}}(K = k|Z_t = z) = \frac{\pi(K = k)\pi(Z_t = z|K = k)}{\pi(Z_t = z)}.$$

Because the domains are convex and disjoint, this shows that $\pi_{\text{mode}}(K = k|Z_t = z) = 1$ if $z \in \Omega_k$ and $\pi_{\text{mode}}(K = k|Z_t = z) = 0$ otherwise. We conclude the proof. □

We now prove Theorem 3.1.

Proof of Theorem 3.1. Let us first prove the straightness bound.

$$u_t(z) = \mathbb{E}_{(Z_0,Z_1)\sim\pi_{\text{mode}}}[Z_1 - Z_0 | Z_t = z]$$

denote the velocity field of the flow between p and q under the mode coupling, and for each $k \in [m]$ and each $t \in [0,1]$, let $Z_t^{(k)} = \mathbb{E}[Z_t | Z_t \in \Omega_k]$. Then, using Lemma 6.1 to factorize the mode coupling, we have

$$\begin{aligned}
\text{Straightness}(p, q; \pi_{\text{mode}}) &= \mathbb{E}\left[\int_0^1 \|u_t(Z_t)\|^2 dt - \left\|\int_0^1 u_t(Z_t)dt\right\|^2\right] \\
&= \mathbb{E}_K \mathbb{E}_{Z_t^{(K)}}\left[\int_0^1 \|u_t(Z_t^{(K)})\|^2 dt - \left\|\int_0^1 u_t(Z_t^{(K)})dt\right\|^2 \Big| K\right] \\
&= \sum_{k=1}^m c_k \mathbb{E}_{Z_t^{(k)}}\left[\int_0^1 \|u_t(Z_t^{(k)})\|^2 dt - \left\|\int_0^1 u_t(Z_t^{(k)})dt\right\|^2 \Big| K = k\right],
\end{aligned}$$

where we used that $\pi_{\text{mode}}(Z_t \in \Omega_k) = c_k$ according to Lemma 6.1. By Lemma 6.2, conditioned on the value of $K = k$, we have we have $u_t(Z_t^{(k)}) = u_t^{(k)}(Z_t^{(k)})$, where $u_t^{(k)}$ is the velocity field that drives the flow between p_k and q_k under the independent coupling; hence the law of the flow $\{Z_t^{(k)}\}_{t\in(0,1)}$ coincides with the law of the flow driven by the velocity field $u_t^{(k)}$ Hence, it follows that, for $k \in [m]$,

$$\mathbb{E}_{Z_t^{(k)}}\left[\int_0^1 \|u_t(Z_t^{(k)})\|^2 dt - \left\|\int_0^1 u_t(Z_t^{(k)})dt\right\|^2 \Big| K = k\right] = \text{Straightness}(p_k, q_k; \pi_{\text{ind}}).$$

This proves the equality in Theorem 3.1. To prove the inequality, note that by Lemma 6.4, we have

$$\begin{aligned}
\text{Straightness}(p_k, q_k; \pi_{\text{ind}}) &= \mathbb{E}_{\tilde{Z}_t}\left[\int_0^1 \|\tilde{u}_t(\tilde{Z}_t)\|_{\Sigma_k}^2 dt - \left\|\int_0^1 \tilde{u}_t(\tilde{Z}_t)dt\right\|_{\Sigma_k}^2\right] \\
&= \mathbb{E}\left[\int_{(t,s)\in(0,1)^2} \|\tilde{u}_t(\tilde{Z}_t) - \tilde{u}_s(Z_s)\|_{\Sigma_k}^2 ds dt\right],
\end{aligned}$$

where $\tilde{u}_t$ is the velocity field of the flow between $\tilde{p}$ and $\tilde{q}$ under the independent coupling. Note that while Lemma 6.4 proves the result with the norm $\|\cdot\|_{\Sigma_k}$ replaced by the Euclidean norm, a parallel argument proves the inequality above (i.e., the choice of norm is unimportant). It therefore follows that

$$\begin{aligned}
\text{Straightness}(p, q; \pi_{\text{mode}}) &= \sum_{k=1}^m c_k \mathbb{E}_{Z_t^{(k)}}\left[\int_0^1 \|u_t(Z_t^{(k)})\|^2 dt - \left\|\int_0^1 u_t(Z_t^{(k)})dt\right\|^2 \Big| K = k\right] \\
&= \sum_{k=1}^m c_k \mathbb{E}_{\tilde{Z}_t}\left[\int_0^1 \|\tilde{u}_t(\tilde{Z}_t)\|_{\Sigma_k}^2 dt - \left\|\int_0^1 \tilde{u}_t(\tilde{Z}_t)dt\right\|_{\Sigma_k}^2\right] \\
&= \sum_{k=1}^m \mathbb{E}\left[\int_{(t,s)\in(0,1)^2} \|\tilde{u}_t(\tilde{Z}_t) - \tilde{u}_s(Z_s)\|_{\Sigma_k}^2 ds dt\right] \\
&\le \sum_{k=1}^m c_k \|\Sigma_k\|_{\text{op}} \mathbb{E}\left[\int_{(t,s)\in(0,1)^2} \|\tilde{u}_t(\tilde{Z}_t) - \tilde{u}_s(Z_s)\|^2 ds dt\right] \\
&= \left(\sum_{k=1}^m c_k \|\Sigma_k\|_{\text{op}}\right) \text{Straightness}(\tilde{p}, \tilde{q}; \pi_{\text{ind}}),
\end{aligned}$$

where we used Lemma 6.4 again to rewrite the straightness of $(\tilde{p}, \tilde{q}; \pi_{\text{ind}})$. To prove the inequality for total length, a parallel argument to the one used to prove the straightness bound shows that

$$\begin{aligned}
\text{Len}(p, q; \pi_{\text{mode}}) &= \mathbb{E} \int_0^1 \|u_t(Z_t)\| dt \\
&= \sum_{k=1}^m c_k \mathbb{E}_{Z_t^{(k)}} \left[\int_0^1 \|u_t(Z_t^{(k)})\| dt \right] \\
&= \sum_{k=1}^m c_k \mathbb{E}_{\tilde{Z}_t} \left[\int_0^1 \|\tilde{u}_t(\tilde{Z}_t)\|_{\Sigma_k} dt \right] \\
&\leq \left(\sum_{k=1}^m c_k \|\Sigma_k^{1/2}\|_{\text{op}} \right) \mathbb{E}_{\tilde{Z}_t} \left[\int_0^1 \|\tilde{u}_t(\tilde{Z}_t)\|_{\Sigma_k} dt \right] \\
&= \left(\sum_{k=1}^m c_k \|\Sigma_k^{1/2}\|_{\text{op}} \right) \text{Len}(\tilde{p}, \tilde{q}; \pi_{\text{ind}}).
\end{aligned}$$

To prove the Lipschitz constant bound, recall from Lemma 6.2 that the velocity field u_t is defined on the union $\bigcup_{k=1}^m \Omega_k$, and that for $z \in \Omega_k$, $u_t(z) = u_t^{(k)}(z)$, where $u_t^{(k)}$ is the velocity field bridging p_k and q_k under the independent coupling. Additionally, since p_k and q_k are images of $\tilde{p}$ and $\tilde{q}$ by the affine map $z \mapsto \Sigma^{1/2} z + \mu$, Lemma 6.3 guarantees that for all $z \in \Omega_k$,

$$u_t^{(k)}(z) = \Sigma^{1/2} \tilde{u}_t(\Sigma_k^{-1/2}(z - \mu_k)),$$

where $\tilde{u}_t$ is the velocity field associated to the flow between $\tilde{p}$ and $\tilde{q}$ under the independent coupling. By Assumption 1, $\tilde{u}_t$ is Lipschitz continuous and its Jacobian is defined globally on Ω. It follows that, for $z \in \Omega_k$,

$$\begin{aligned}
\|\nabla u_t(z)\|_{\text{op}} &= \left\| \nabla \left(\Sigma_k^{1/2} \tilde{u}_t(\Sigma_k^{-1/2}(z - \mu_k)) \right) \right\|_{\text{op}} \\
&= \left\| \Sigma_k^{1/2} \left(\nabla \tilde{u}_t \right) (\Sigma_k^{-1/2}(z - \mu_k)) \Sigma_k^{-1/2} \right\|_{\text{op}} \\
&= \left\| \left(\nabla \tilde{u}_t \right) (\Sigma_k^{-1/2}(z - \mu_k)) \right\|_{\text{op}} \\
&\leq \text{Lip}(\tilde{u}_t).
\end{aligned}$$

Hence it follows that

$$\text{Lip}(u_t) = \sup_{z \in \bigcup_k \Omega_k} \|\nabla u_t(z)\|_{\text{op}} \leq \text{Lip}(\tilde{u}_t).$$

We conclude the proof. □

Proof of Theorem 3.2. In this example, we take $k = 2$ and set $\mu_1 = 0$, $\mu_2 = \mu$, $\Sigma_1 = I$, and $\Sigma_2 = \sigma^2 I$. Define $\Omega' = \{\sigma x + \mu : x \in \Omega\}$, and assume that μ and σ are chosen so that Ω' is at a positive distance from Ω. We consider the velocity field $u_t(x)$ arising from the flow between the unimodal prior $\tilde{p}$ and the target $q(x) = \frac{1}{2}(\tilde{q}(x) + \tilde{q}(\sigma^{-1}(x - \mu)))$. Observe that the velocity field $u_t(x)$ bridging $\tilde{p}$ and q under the independent coupling is given by

$$\begin{aligned}
u_t(x) &= \mathbb{E}[Z_1 - Z_0 | Z_t = x] \\
&= \mathbb{P}(Z_1 \in \Omega | Z_t = x) \tilde{u}_t(x) + \mathbb{P}(Z_1 \in \Omega' | Z_t = x) u_t^{\mu,\sigma}(x),
\end{aligned}$$

where $\tilde{u}_t(x)$ is the velocity field bridging $\tilde{p}$ and $\tilde{q}$ under the independent coupling and $u_t^{\mu,\sigma}$ is the velocity field bridging $\tilde{p}$ and $\tilde{q}(\sigma^{-1}(x - \mu))$. Note that since Ω' is at a positive distance from Ω, the sets $\text{Lin}(\Omega, \Omega') = \{t z_1 + (1 - t) z_0 : z_0 \in \Omega,\ z_1 \in \Omega',\ t \in (0, 1)\}$ and $\text{Lin}(\Omega, \Omega) = \{t z_1 + (1 - t) z_0 : z_0, z_1 \in \Omega\ t \in (0, 1)\} \subseteq \Omega$ are disjoint. It follows that the posterior probabilities $\mathbb{P}(Z_1 \in \Omega | Z_t = z)$ and $\mathbb{P}(Z_1 \in \Omega' | Z_t = z)$ are $\{0, 1\}$-valued, and hence the velocity field $u_t(z)$ can be expressed as

$$u_t(z) = \begin{cases} \tilde{u}_t(z), & z \in \text{Lin}(\Omega, \Omega) \\ u_t^{\mu,\sigma}(z), & z \in \text{Lin}(\Omega, \Omega'). \end{cases}$$

Therefore, the Lipschitz constant of u_t satisfies

$$\begin{aligned}\mathrm{Lip}(u_t) &= \sup_z \|\nabla u_t(z)\|_{\mathrm{op}} \\ &\geq \sup_{z \in \mathrm{Lin}(\Omega,\Omega')} \|\nabla u_t(z)\|_{\mathrm{op}} \\ &= \sup_z \|\nabla u_t^{\mu,\sigma}(z)\|_{\mathrm{op}} \\ &= \mathrm{Lip}(u_t^{\mu,\sigma}).\end{aligned}$$

This shows that it suffices to lower bound the Lipschitz constant of the velocity field between $\tilde{p}$ and $\tilde{q}(\sigma^{-1}(z-\mu))$. In addition, observe that the velocity fields $u_t^{\mu,\sigma}$ and $u_t^{0,\sigma}$ (i.e., the latter is obtained by setting $\mu = 0$) are related by the formula

$$u_t^{\mu,\sigma}(z) = \mu + u_t^{0,\sigma}(z - t\mu).$$

This shows that the Lipschitz constant of the velocity field is independent of the choice of μ, so we take $\mu = 0$ without loss of generality. In slight abuse of notation, we abbreviate $u_t^{0,\sigma}(z)$ to $u_t(z)$ and define $L_t = \mathrm{Lip}(u_t)$. Our goal is to lower bound the time-averaged Lipschitz constant $\int_0^1 L_t dt$ by an increasing function of the variance σ^2.

Before proceeding with the proof, we introduce some more notation and assumptions. We let $\Phi_t(\cdot)$ denote the flow map and $\Phi_t^{-1}(\cdot)$ the inverse of the flow map, and we make the following assumptions:

1. For each (t, z), the Jacobian of Φ_t is invertible;
2. Φ_1 and Φ_1^{-1} are both Lipschitz continuous.

We believe the results can be generalized without the assumptions above, but we adopt them to allow for the most simple proof (for a discussion of the Lipschitz continuity of flow maps associated to flow matching, see [16]). We denote by $M = \sup_z \|\nabla(\Phi_1^{-1}(z))\|_{\mathrm{op}}$ the Lipschitz constant of Φ_1^{-1} and write $L_t := \sup_z \|\nabla u_t(z)\|_{\mathrm{op}}$ for the Lipschitz constant of u_t at time t.

For the proof, we proceed as follows. Define the function $B_t(z) = \nabla(\Phi_t^{-1}(\Phi_t(z)))$. By Lemma 6.5, for each $z \in \mathbb{R}^d$, $B_t(z)$ solves the ODE

$$\frac{d}{dt} B_t(z) = -B_t(z)\nabla u_t(z).$$

Using the chain rule, the Cauchy-Schwarz inequality, and the inequality $\|AB\|_F \leq \|A\|_F \|B\|_{\mathrm{op}}$, we deduce that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|B_t(z)\|_F^2 &= \langle B_t(z), \frac{d}{dt} B_t(z)\rangle_F \\ &= \langle B_t(z), B_t(z)\nabla u_t(z)\rangle_F \\ &\leq \|B_t(z)\|_F^2 \|\nabla u_t(z)\|_{\mathrm{op}} \\ &\leq L_t \|B_t(z)\|_F^2.\end{aligned}$$

By Grönwall's inequality, we therefore have

$$\begin{aligned}\|B_1(z)\|_F^2 &\leq \|B_0(z)\|_F^2 \exp\left(2\int_0^1 L_t dt\right) \\ &= d\exp\left(2\int_0^1 L_t dt\right),\end{aligned}$$

where we used that $B_0(z) = I$. Since the above inequality holds for any z, it follows that

$$\begin{aligned}M &:= \sup_z \|\nabla(\Phi_1^{-1})(\Phi_1(z))\|_{\mathrm{op}} && (6.1)\\ &= \sup_z \|B_t(z)\|_{\mathrm{op}} && (6.2)\\ &\leq \sup_z \|B_t(z)\|_F && (6.3)\\ &\leq \sqrt{d}\exp\left(\int_0^1 L_t dt\right). && (6.4)\end{aligned}$$

Having deduced an upper bound on M, we now find a corresponding lower bound in terms of the variance σ^2. Notice that under our assumptions, the inverse function theorem implies that $\nabla(\Phi_1^{-1})(\Phi_1(z)) = (\nabla\Phi_1(z))^{-1}$, where the latter inverse denotes the matrix inverse of the Jacobian. Hence, M being finite implies that, for any z, y, $\|\Phi_1(z) - \Phi_1(y)\| \geq \frac{1}{M}\|z - y\|$. Since Φ_1 is a transport map between $\tilde{p}$ and q, this implies that

$$\begin{aligned}
d\sigma^2 &= \text{Trace}(\text{Cov}(q)) \\
&= \frac{1}{2}\mathbb{E}_{X,Y\sim q, X\perp Y}[\|X - Y\|^2] \\
&= \frac{1}{2}\mathbb{E}_{X,Y\sim \tilde{p}, X\perp Y}[\|\Phi_1(X) - \Phi_1(Y)\|] \\
&\geq \frac{M^2}{2}\mathbb{E}_{X,Y\sim \tilde{p}, X\perp Y}[\|X - Y\|^2] \\
&= \frac{1}{M^2}\text{Trace}(\text{Cov}(p)) \\
&= \frac{d}{M^2},
\end{aligned}$$

which implies that $M \geq \sigma^{-1}$. Combining this with inequality 6.1, we have that

$$\begin{aligned}
&\sigma^{-1} \leq \sqrt{d}\exp\left(\int_0^1 L_t dt\right) \\
\implies &\log(\sigma^{-1}) \leq \frac{1}{2}\log(d) + \int_0^1 L_t dt \\
\implies &\int_0^1 L_t dt \geq \log(\sigma^{-1}) - \frac{1}{2}\log(d),
\end{aligned}$$

and hence $\int_0^1 L_t dt$ can be made arbitrarily large by choosing σ accordingly. We conclude the proof. □

6.1. Auxiliary lemmas

We use the following lemmas at various points.

Lemma 6.3. *[Velocity field under affine transforms] Let p and q be probability distributions and let $u_t(z) = \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[Z_1 - Z_0|Z_t = z]$ denote their velocity field. Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}_{sym}$. Let p_μ, q_μ denote the pushforward distributions of p and q along the map $z \mapsto z + \mu$, and let p_Σ, q_Σ denote the pushforward distributions of p and q along the map $z \mapsto \Sigma^{1/2}z$.*

1. *The velocity field between p_μ and q_μ is given by*

$$(t, z) \mapsto u_t(z - \mu),$$

   for all $t \in (0, 1)$ and all $z \in \mathbb{R}^d$ for which $u_t(z - \mu)$ is defined.

2. *The velocity field between p_Σ and q_Σ is given by*

$$(t, z) \mapsto \Sigma^{1/2}u_t(\Sigma^{-1/2}z),$$

   for all $t \in (0, 1)$ and $z \in \mathbb{R}^d$ for which $u_t(\Sigma^{-1/2}z)$ is defined.

Proof. For 1), the velocity field is given by

$$\begin{aligned}
(t, z) &\mapsto \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\mu\otimes q_\mu}[\tilde{Z}_1 - \tilde{Z}_0|(1 - t)\tilde{Z}_0 + t\tilde{Z}_1 = z] \\
&= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[Z_1 - Z_0|(1 - t)Z_0 + tZ_1 = z - \mu] \\
&= u_t(z - \mu).
\end{aligned}$$

For 2), the velocity field is given by

$$(t, z) \mapsto \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\mu\otimes q_\mu}[\tilde{Z}_1 - \tilde{Z}_0|(1 - t)\tilde{Z}_0 + t\tilde{Z}_1 = z]$$

$$
\begin{aligned}
&= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[\Sigma^{1/2}(Z_1 - Z_0)|\Sigma^{1/2}((1-t)Z_0 + tZ_1) = z] \\
&= \Sigma^{1/2}\mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[Z_1 - Z_0|(1-t)Z_0 + tZ_1 = \Sigma^{-1/2}z] \\
&= \Sigma^{1/2}u_t(\Sigma^{-1/2}z).
\end{aligned}
$$

□

Lemma 6.4. *[Straightness under affine transforms] Assume the notation of Lemma 6.3. Then we have*

$$\textit{Straightness}(p_\mu, q_\mu) = \textit{Straightness}(p, q).$$

We also have

$$\textit{Straightness}(p_\Sigma, q_\Sigma) = \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\int_0^1 u_t(Z_t)dt\right\|_\Sigma^2\right] - \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\int_0^1 u_t(Z_t)dt\right\|_\Sigma^2\right].$$

Lemma 6.4 states that the straightness is preserved under translation. It also states that the straightness of p_Σ and q_Σ is equal to the straightness of p and q, but where distance is measured by the norm $\|\cdot\|_\Sigma$ instead of the Euclidean norm.

Proof. Given two distributions p, q, we write their straightness as a sum of two terms, namely

$$\text{Straightness}(p, q) = S_1(p, q) + S_2(p, q),$$

where

$$
\begin{aligned}
S_1(p, q) &= \int_0^1 \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[\|u_t(Z_t)\|^2]dt \\
S_2(p, q) &= \mathbb{E}_{Z_t}\left[\left\|\int_0^1 u_t(Z_t)dt\right\|^2\right].
\end{aligned}
$$

To prove the results for (p_μ, q_μ) let u_t^μ denote the velocity field between p_μ and q_μ. Then we have

$$
\begin{aligned}
S_1(p_\mu, q_\mu) &= \int_0^1 \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\mu\otimes q_\mu}\|u_t^\mu(\tilde{Z}_t)\|^2 dt \\
&= \int_0^1 \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\mu\otimes q_\mu}[\|\mathbb{E}[\tilde{Z}_1 - \tilde{Z}_0|\tilde{Z}_t\|^2] \\
&= \int_0^1 \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[\|\mathbb{E}[Z_1 - Z_0|Z_t + \mu]\|^2] \\
&= \int_0^1 \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\|u_t(Z_t)\|^2 dt \\
&= S_1(p, q),
\end{aligned}
$$

where we used the equality $u_t^\mu(z) = u_t(z - \mu)$ derived in Lemma 6.3. Similarly, for S_2, we have

$$
\begin{aligned}
S_2(p_\mu, q_\mu) &= \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\mu\otimes q_\mu}\left[\left\|\int_0^1 u_t^\mu(\tilde{Z}_t)dt\right\|^2\right] \\
&= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\int_0^1 u_t^\mu(Z_t + \mu)dt\right\|^2\right] \\
&= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\int_0^1 u_t(Z_t)dt\right\|^2\right] \\
&= S_2(p, q).
\end{aligned}
$$

To prove the results for (p_Σ, q_Σ), let u_t^Σ denote the velocity field between p_Σ and q_Σ. We have

$$\begin{aligned} S_1(p_\Sigma, q_\Sigma) &= \int_0^1 \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\Sigma\otimes q_\Sigma}[\|u_t^\Sigma(\tilde{Z}_t)\|^2]dt \\ &= \int_0^1 \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\Sigma\otimes q_\Sigma}[\|\Sigma^{1/2}u_t(\Sigma^{-1/2}\tilde{Z}_t)\|^2]dt \\ &= \int_0^1 \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[\Sigma^{1/2}\|u_t(Z_t)\|^2]dt \\ &= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}[\|u_t(Z_t)\|_\Sigma^2]dt \end{aligned}$$

and

$$\begin{aligned} S_2(p_\Sigma, q_\Sigma) &= \mathbb{E}_{(\tilde{Z}_0,\tilde{Z}_1)\sim p_\Sigma\otimes q_\Sigma}\left[\left\|\int_0^1 u_t^\Sigma(\tilde{Z}_t)\right\|^2\right] \\ &= \mathbb{E}_{\tilde{Z}_0,\tilde{Z}_1}\left[\left\|\int_0^1 \Sigma^{1/2}u_t(\Sigma^{-1/2}\tilde{Z}_t)\right\|^2\right] \\ &= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\Sigma^{1/2}\int_0^1 u_t(Z_t)dt\right\|^2\right] \\ &= \mathbb{E}_{(Z_0,Z_1)\sim p\otimes q}\left[\left\|\int_0^1 u_t(Z_t)dt\right\|_\Sigma^2\right]. \end{aligned}$$

We conclude the proof. □

The next Lemma describes how the Jacobian of the flow map evolves in time.

Lemma 6.5. *Let $u : (0,1)\times\mathbb{R}^d \to \mathbb{R}^d$ be a velocity field and let $\Phi : (0,1)\times\mathbb{R}^d \to \mathbb{R}^d$ be the flow map associated to the ODE*

$$\begin{cases} \frac{d}{dt}\Phi_t(z) = u_t(\Phi_t(z)), t\in[0,1) \\ \Phi_0(z) = z. \end{cases}$$

Define $A_t(z) = \nabla\Phi_t(z)$ to be the Jacobian of the flow map and $B_t(z) = \nabla(\Phi_t^{-1}(z))$ to be the Jacobian of the inverse of the flow map. Assume that the matrix $A_t(z)$ is invertible for each (t,z). Then A_t and B_t are related by the Equation

$$A_t(z)B_t(\Phi_t(z)) = B_t(\Phi_t(z))A_t(z) = I,$$

and for each z, $A_t(z)$ and $B_t(z)$ satisfy the ODEs

$$\begin{cases} \frac{d}{dt}A_t(z) = \nabla u_t(\Phi_t(z))A_t(z),\ t\in[0,1) \\ A_0(z) = I_d, \end{cases}$$

and

$$\begin{cases} \frac{d}{dt}B_t(\Phi_t(z)) = -B_t(\Phi_t(z))\nabla u_t(z),\ t\in[0,1) \\ B_0(\Phi_0(z))) = I \end{cases}.$$

Proof. Let us begin by deriving the dynamics of $A_t(z)$. By the chain rule, we have

$$\begin{aligned} \frac{d}{dt}A_t(z) &= \frac{d}{dt}\nabla\Phi_t(z) \\ &= \nabla\frac{d}{dt}\Phi_t(z) \end{aligned}$$

$$
\begin{aligned}
&= \nabla(u_t \circ \Phi_t)(z) \\
&= \nabla u_t(\Phi_t(z)) \nabla \Phi_t(z) \\
&= \nabla u_t(\Phi_t(z)) A_t(z).
\end{aligned}
$$

To derive the relationship between A_t and B_t, note that the inverse function theorem and the invertibility of $A_t(z)$ imply that

$$
\begin{aligned}
B_t(\Phi_t(z))) &= (\nabla \Phi_t^{-1})(\Phi_t(z)) \\
&= (\nabla \Phi_t(z))^{-1} \\
&= A_t(z)^{-1},
\end{aligned}
$$

where the outer inverse denotes the matrix inverse. We can use this to derive the dynamics of $B_t(z)$: using the product rule, we have

$$
\begin{aligned}
0 = \frac{d}{dt}(A_t(z) B_t(\Phi_t(z))) &= \frac{d}{dt}(A_t(z)) B_t(\Phi_t(z)) + A_t(z) \frac{d}{dt} B_t(\Phi_t(z)) \\
\implies -\nabla u_t(\Phi_t(z)) \underbrace{A_t(z) B_t(\Phi_t(z))}_{=I} &= A_t(z) \frac{d}{dt} B_t(\Phi_t(z)) \\
\implies -u_t(\Phi_t(z)) &= A_t(z) \frac{d}{dt} B_t(\Phi_t(z)) \\
\implies \frac{d}{dt} B_t(z) &= -A_t(z)^{-1} \nabla u_t(\Phi_t(z)) \\
&= -B_t(\Phi_t(z)) \nabla u_t(\Phi_t(z)).
\end{aligned}
$$

□

7. Toy Experiments

MM−FM as the source distribution has some preferred trajectory properties compared to unimodal Gaussian when the target distribution is multi-modal, such as shorter and straighter trajectories. We design a series of low-dimensional simulations to reveal them.

7.1. Experimental Setup

Target distribution. We define the target distributions as GMMs with 100 components in a 10-dimensional space ($d = 10$), denoted by $f(x) = \sum_{i=1}^{100} c_i N(x; \mu_i, \Sigma_i)$. The mean vectors μ_i are independently drawn from an isotropic Gaussian distribution $N(0, \sigma^2 I)$. The covariance matrix of each component is independently generated by $\Sigma_i = \frac{A_i^T A_i}{d+1}$, where A_i is a $d \times d$ matrix with each entry sampled independently from the standard normal distribution $N(0, 1)$. The mixture weights are drawn from a symmetric Dirichlet distribution Dir(α), where all concentration parameters are set to the dimensionality of the space (i.e. $\alpha_k = d$ for all components k). To create distinct simulation scenarios, we vary the scale of the means by setting σ to 0.7, 1.0, and 2.0, which we refer to as the *compact*, *normal*, and *spread* settings respectively. For each setting, we generate a dataset of 80,000 points, which is then standardized to curate the training set.

Model and training. The network architecture in all experiments is a Multilayer Perceptron (MLP) with 4 hidden layers, each containing 256 neurons and using the ReLU activation function. All setups utilize a linear transport path, and the MLP is trained to model the underlying dynamics. For models trained with a GMM prior, the number of components is set to 6. We optimize the model by minimizing the Mean Squared Error (MSE) loss using the AdamW optimizer with a learning rate of 5×10^{-4} and a weight decay of 10^{-4}. The model is trained for 800 epochs with a batch size of 2048. This entire experimental setup remains identical across the three data configurations and all methods to ensure a fair comparison.

Evaluation. We evaluate the performance the basis of two primary aspects: the quality of the generated samples and the properties of the learned trajectories. To assess sample quality, we employ the `geomloss` package to compute the 2-Wasserstein Distance of generated samples to 10,000 held-out true samples that is approximated using Sinkhorn algorithm with an entropic regularization of $\epsilon = 0.05$. The scores for each metric are calculated over 10 independent sampling runs

Data Setup	Model Prior	Coupling	2-Wasserstein(↓)	Length(↓)	Straightness (↓)
Compact	Gaussian	Independent	1.252 ± 0.002	1.876 ± 0.5091	4.067 ± 2.126
	GMM	Independent	1.251 ± 0.002	1.820 ± 0.5234	4.070 ± 2.227
		Mode Coupling	$\mathbf{1.248 \pm 0.003}$	$\mathbf{1.665 \pm 0.4788}$	$\mathbf{3.258 \pm 1.829}$
Normal	Gaussian	Independent	1.366 ± 0.001	1.972 ± 0.4912	4.073 ± 1.880
	GMM	Independent	1.368 ± 0.003	1.921 ± 0.4972	4.148 ± 2.016
		Mode Coupling	$\mathbf{1.364 \pm 0.003}$	$\mathbf{1.717 \pm 0.4564}$	$\mathbf{3.094 \pm 1.593}$
Spread	Gaussian	Independent	1.582 ± 0.006	2.346 ± 0.5212	4.245 ± 1.664
	GMM	Independent	1.631 ± 0.006	2.300 ± 0.5609	4.651 ± 2.225
		Mode Coupling	$\mathbf{1.556 \pm 0.003}$	$\mathbf{1.999 \pm 0.4816}$	$\mathbf{3.002 \pm 1.299}$

Table 5. The GMM source distribution with mode coupling leads to better sample quality, straighter and shorter trajectories in all compact, normal, and spread multimodal target distributions.

and then averaged. Given a trajectory $\{\psi_t(X_0)\}_{t=0}^1$ starting from X_0 with an underlying velocity field u_t, the length and straightness are computed by numerically computing the following integrals:

$$\text{Length}(\{\psi_t(X_0)\}_{t=0}^1) = \int_0^1 \|u_t(\psi_t(X_0)\| dt$$

$$\begin{aligned}\text{Straightness}(\{\psi_t(X_0)\}_{t=0}^1) &= \int_0^1 \|u_t(\psi_t(X_0))\|^2 dt - \left\| \int_0^1 u_t(\psi_t(X_0)) dt \right\|^2 \\ &= \int_0^1 \|u_t(\psi_t(X_0))\|^2 dt - \|\psi_1(X_0) - X_0\|^2\end{aligned}$$

The trajectory metrics are computed and averaged over 10,000 starting points from a single sampling procedure. We use the Euler solver with 40 steps for all ODE solving.

7.2. Results and Analysis

The quantitative results are summarized in Tab. 5. Regarding sample quality, models trained with a GMM prior and equipped with mode coupling consistently outperform those trained with an isotropic Gaussian prior or with independent coupling under all data configurations. We further track the 2-Wasserstein distance during training at intervals of 50 epochs, as illustrated in Figure 5, to demonstrate that our model achieves better sample quality throughout the training process.

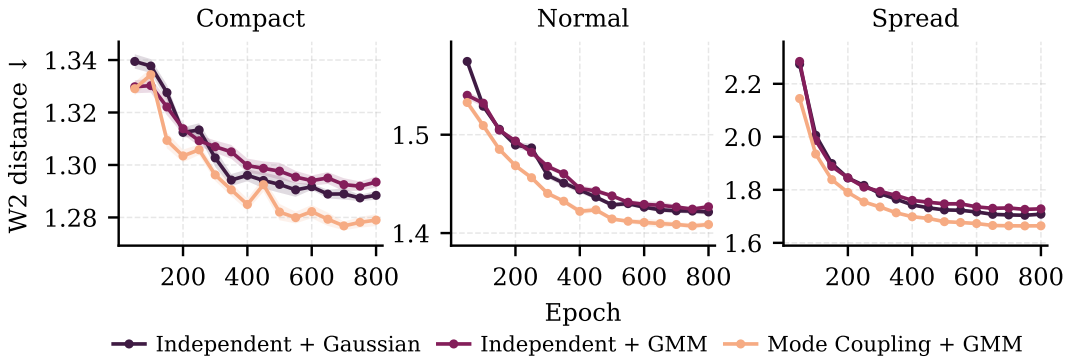

Figure 5. The 2-Wasserstein distance along training. `MM-FM` consistently leads to faster training convergence and better sample quality.

For the trajectory metrics, model trained with GMM prior with non-trivial coupling also outperforms the independent GMM and Isotropic Gaussian prior, in both trajectory length and straightness. As the modes become more well-separated moving from *compact* to *spread* setup, the improvement is increasingly evident. This suggests that our method can benefit from a target distribution with more separable modes, explaining the significance of the choice of DINOv2-B as tokenizer in our ImageNet 256×256 experiment.

7.3. Boosting OT-CFM with MM-FM

BatchOT[64] and OT-CFM[77] are two similar frameworks that leverage optimal transport (OT) to define couplings between data and noise samples at the minibatch level. Their methods aims to approximate the dynamic OT flow that bridges the target and source distribution and thus straightening the flow. We propose that the performance of their methods can be improved further when combined with MM−FM.

According to OT-CFM, the performance of model trained with minibatch OT benifits from larger batch size, since the minibatch OT plan obtained from a larger batch better approximates the data-level OT plan. But in practice the algorithm is prohibited from using large batch size, because of the $\mathcal{O}(n^3)$ time complexity of OT algorithm. One clear advantage of MM−FM is that the mass of GMM prior in local regions better matches the target distribution, enabling moving mass locally to shift to the target distribution, and mode coupling reduces unnecessary coupling of distant point pairs, leading to a more batch-size-efficient way of solving minibatch OT. We explore the probability of injecting such prior information into OT-CFM and found that the trained model using the combined method yields straighter trajectory than that trained using OT-CFM alone.

We combine OT-CFM and MM−FM in the following manner: In each iteration, given a batch of size b from the target distribution $\{X_1^{(i)}\}_{i=1}^b$ and the GMM prior $p(x) = \sum_{k=1}^m c_k \mathcal{N}(x; \mu_k, \sigma_k^2 I)$, we first assign each point $X_1^{(i)}$ to their respective mode $k^{(i)}$ using the GMM prior, and then sample the $X_0^{(i)}$ from the mode $k^{(i)}$ using $X_0^{(i)} \sim \mathcal{N}(\mu_{k^{(i)}}, \sigma_{k^{(i)}}^2 I)$, as we did when training MM−FM. Instead of computing OT over the whole batch, we perform OT for all points within the same mode. Samples are drawn from the mode-specific OT plan for each mode and then concatenated to constitute the training set for this batch. Then we draw X_t and perform gradient descend in the same manner as we did in MM−FM. We refer to this combined coupling scheme as Mode+OT-CFM and the detailed training algorithm is shown in Algorithm 3.

Algorithm 3 Training OT-CFM combined with MM−FM

Input: samples $\{x_1^{(j)}\}$ from multimodal distribution q, batch size b
Output: a velocity field neural network $u_t^\theta(X_t)$

Stage 1: Gaussian Mixture Model (GMM) Fitting
$m \leftarrow$ Bayesian GMM($\{x_1^{(j)}\}$) ▷ *Infer number of modes*
Fit GMM $p = \sum_{i=1}^m c_i \mathcal{N}(\mu_i, \Sigma_i)$ to $\{x_1^{(j)}\}$

Stage 2: Flow Matching Training
Initialize parameters θ for $u_t^\theta(X_t)$
for each training iteration **do**
  Sample $(X_0, X_1) \sim \pi_{\text{mode}}(x_0|x_1)$ via mode coupling:
    Sample $\{X_1^{(i)}\}_{i=1}^b \overset{i.i.d}{\sim} q$,
    Infer mode $k^{(i)} \leftarrow \arg\max_j c_j \mathcal{N}(X_1^{(i)}; \mu_j, \Sigma_j)$
    Sample $X_0^{(i)} \sim \mathcal{N}(\mu_{k^{(i)}}, \Sigma_{k^{(i)}})$
  Solve minibatch level OT within each mode:
    for $j = 1$ **to** m **do**
      $X^{(i)} = \{(X_0^{(i)}, X_1^{(i)}) \mid k^{(i)} = j\}$
      $\pi^{(j)} \leftarrow \text{OT}(X^{(j)})$
      Sample $X^{(j)'} = \{(X_0^{(i)'}, X_1^{(i)'})\}_{i=1}^{|X^{(j)}|} \sim \pi^{(j)}$
    end for
  $X_0 = \left\{X_0^{(j)'}\right\}_{j=1}^m, X_1 = \left\{X_1^{(j)'}\right\}_{j=1}^m$
  Sample $t \sim \mathcal{U}(0,1)$ and set $X_t = (1-t)X_0 + tX_1$
  Descend gradient on loss $\mathcal{L}_{\text{CFM}}^{OT}(\theta)$:
    $\nabla_\theta \|u_t^\theta(X_t) - (X_1 - X_0)\|^2$

To empirically verify the effectiveness of the combined coupling method, we perform an experiment with the identical setting as is introduced in Sec. 7.1. The only modification we made is on the batch size b used in training: while in the two other competing methods (Gaussian prior and GMM prior with OT-CFM coupling) we employed $b = 256$, we increased the batch size used in Mode+OT-CFM to $b = 1024$. Note that since we divided the OT problem to sub-problems within each component, and the number of components of our GMM prior is 6, the average effective scale of OT problem solved is on

Data Setup	Model Prior	Coupling	2-Wasserstein(↓)	Length(↓)	Straightness (↓)	OT Size(↓)
Compact	Gaussian	OT-CFM	$\mathbf{1.248 \pm 0.002}$	0.7428 ± 0.3621	0.3150 ± 0.4228	256
	GMM	OT-CFM	1.252 ± 0.003	0.6257 ± 0.3368	0.3000 ± 0.3814	256
		Mode+OT-CFM	1.253 ± 0.003	$\mathbf{0.6253 \pm 0.3107}$	$\mathbf{0.2507 \pm 0.2981}$	170
Normal	Gaussian	OT-CFM	$\mathbf{1.362 \pm 0.001}$	0.8993 ± 0.3879	0.3028 ± 0.3493	256
	GMM	OT-CFM	1.367 ± 0.003	0.7701 ± 0.3528	0.2908 ± 0.3240	256
		Mode+OT-CFM	1.366 ± 0.003	$\mathbf{0.7532 \pm 0.3293}$	$\mathbf{0.2240 \pm 0.2429}$	170
Spread	Gaussian	OT-CFM	1.559 ± 0.005	1.472 ± 0.4466	0.3207 ± 0.2406	256
	GMM	OT-CFM	1.552 ± 0.004	1.205 ± 0.3915	0.2952 ± 0.2195	256
		Mode+OT-CFM	$\mathbf{1.535 \pm 0.004}$	$\mathbf{1.196 \pm 0.3892}$	$\mathbf{0.1947 \pm 0.1443}$	170

Table 6. Our method `MM-FM` is orthogonal and compatible with methods to solve batch-level OT for optimal coupling. Combining OT-CFM with `MM-FM` can further straighten and shorten the trajectories.

number of points = $\frac{1024}{6} \approx 170$, which is much smaller than that in the other two methods since they are solved on batch size equals 256. The results of the experiment is shown in Tab. 6.

The results in Tab. 6 show that even though our coupling method Mode+OT-CFM perform OT algorithm on a smaller scale, we are still able to achieve much shorter and straighter path than the OT-CFM coupling, while keeping the sample quality comparable.

8. Further Comparison with Similar Works

As discussed in Section 2.3, [35, 36] share a similar spirit to `MM-FM` in that they train generative models with GMM source distributions. However, an important distinction is that their trained velocity fields take conditional variables, such as class labels or mode assignments, as additional inputs, and hence their methodology is not an apples-to-apples comparison with truly unconditional generative models. In more detail, these works consider a prior distribution with m conditional variables $\{c_k\}_{k=1}^m$, each indexing a Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$, and a training objective of the form

$$\sum_{k=1}^{M} \mathbb{E}_{t, x_0 \sim \mathcal{N}(\mu_k, \Sigma_k), x_1 \sim q} \left[\left\| u_t^\theta(x_t, c_k) - u_t(x_t | x_1, c_k) \right\|^2 \right]. \tag{8.1}$$

The global minimizer of the objective above is the vector field given by $(x, c_k) \mapsto u_t(x|c_k)$, the velocity field which optimally bridges $\mathcal{N}(\mu_k, \Sigma_k)$ and the conditional distribution $q(x|c_k)$. In [35], the conditions $\{c_k\}_{k=1}^m$ describe class labels, whereas in [36], they describe mode assignments. In contrast, our velocity field is trained on the OT conditional flow matching loss in Equation equation 2.4, where the coupling π is the mode coupling. In contrast, our velocity field is trained on the standard joint flow matching objective where (X_0, X_1) are coupled according to the mode coupling π_{mode}:

$$u_t^\theta \mapsto \mathbb{E}_{t, (X_0, X_1) \sim \pi_{\text{mode}}} \left[\left\| u_t^\theta(X_t) - u_t(X_t | X_1) \right\|^2 \right]. \tag{8.2}$$

When the source distribution is a mixture of Gaussians, and the conditions $\{c_k\}_{k=1}^m$ describe (hard) mode assignments, it is easy to see that the true minimizer $u_t(x)$ of the above objective is given by

$$u_t(x) = \sum_{k=1}^{m} \pi_{\text{mode}}(c_k | X_t = x) u_t(x|c_k).$$

When the supports of the conditional distributions $\{q(\cdot|c_k)\}_{k=1}^m$ are disjoint, and the mixture prior p is a warm start for q, the posterior probabilities $\pi_{\text{mode}}(c_k|X_t = x)$ converge to being hard assignments, i.e. $\pi_{\text{mode}}(c_k|X_t = x) = 1$ if and only if $x = tx_1 + (1-t)x_0$ for some $x_1 \in \text{supp}(q(\cdot|c_k))$ and $x_0 \in \text{supp}(p(\cdot|c_k))$, and thus *the minimizer of the OT conditional matching objective 2.4 with mode coupling approximately coincides with the minimizer of the mode-conditional objective 8.1*. This demonstrates that unconditional generation can achieve comparable results to conditional generation by leveraging co-design of the coupling and source distribution. It also demonstrates the importance of the foundation models as tokenizers, which ensure that the components of the data distribution are separated in the latent space.

9. More Related Works

Diffusion models as flow-based models. Diffusion models (DMs) [34, 73, 74] are an important sub-class of simulation-free flow-based generative models. In essence, DMs are a special case of FMs which (1) perform generation using a stochastic differential equation rather than an ODE, (2) implement a specific affine probability path, and (3) DMs reparameterize time to run from $[0, T]$ with $T \gg 1$, rather than from $[0, 1]$. The training of DMs via score matching is essentially learning a stage-wise denoiser for the target distribution q. We refer the interested reader to [74] for more details on DMs and [53] for more details on DMs from the FM point of view. The language we use in this paper is following the FM convention but our method is applicable to training DMs as well.

Improvement of flow matching algorithm for learning efficiency. In Sec. 2.1, we follow standard flow matching convention that involves the design choice of source distributions, paths and data couplings, but recent works have also proposed new avenues to improve the learning efficiency. Recognizing that path crossing is inevitable, GMFlow [12] and variational RF [29] model the velocity as a distribution instead of a global mean so that the velocity prediction in high-variance regions due to crossing will not be singular.

Continuous visual tokenizers for latent flow-based models in image generation. Modern-day diffusion and FM models are typically trained in the latent space to optimize training efficiency [7, 20, 81]. As SD-VAE [20] and FLUX-VAE [7] are trained on large scale datasets, they are widely adopted as the default visual encoder of continuous latent space [39, 54, 55, 62]. As mapping image data inputs to Gaussian distributions makes no assumptions of the data structure [40], a series of works attempt to incorporate representations from self-supervised learning (SSL) vision foundation models to impose structured latent spaces so that training FM models can be made efficient. These works include but are not limited to VA-VAE [85], MAETok [11], AlignTok [10], RAE [89], VFM-VAE [6], SVGTok [70], FAE [24], RepTok [28], GigaTok [84]. Other advances include token-efficient design such as DCAE [13], equivariance regularization such as EQ-VAE [42] and end-to-end training with generative models such as E2E-VAE [49]. In spirit, our method works with any structured latent distributions that are well-separable in space (i.e., exhibit multimodal structures), which can be verified quantitatively by linear probing results and approximately qualitatively by t-SNE plots. We acknowledge the rich body of literature on discrete visual tokenizers (i.e., codebook-style) but it is outside of the scope of this paper.

Vision foundation models for improving diffusion transformers. Unlike vision foundation models for latent distributions, which directly corresponds to the target distribution of the FM design space, recent advances also leverage these foundation models to improve diffusion transformers. REPA [87] as the seminal work accelerates DiT convergence by aligning its middle block with the features of the foundation models. REPA-E [49] extend REPA to end-to-end VAE joint training with DiT, and iREPA [72] finds spatial structure matters more than global semantics. HASTE [83] resolves the REPA's diminishing returns during long training via a two-phase schedule that switches when the generative student's capacity needs to be allocated in denoising instead of alignment. While REG [86] speeds up convergence further by adding a learnable token to entangle with the [CLS] token from the foundation models, DDT [82] decouples DiT from the encoder-decoder and applies REPA to the encoder output. ReDi [43] generates latents and foundation model features jointly within a DiT. In contrast, our method is orthogonal to this line of work because we modify the source distribution and data coupling, leaving the DiT untouched. Meanwhile, these methods can be compatible and applied directly together with our method, which can be viewed as imposing further regularized learning signals on the DiT, but we leave them for future work.

10. Implementation Details

Encoder Normalization Unlike [89], we do not discard [CLS] token because it help induce an operational GMM for the latent (patch tokens) space (see Fig. 2). We find the [CLS] token from DINOv2 with Registers [17] are already normalized. For patch tokens, we apply normalization using the pre-computed statistics by [89].

10.1. Training GMM

For ImageNet256 dataset, we train GMM using `scikit-learn` on the [CLS] token array of shape [1281167, 768] because training on patch tokens array of shape [1281167, 196608] suffers from curse of dimensionality and instable likelihood estimation in high dimensions. We configure the GMM with the following hyperparameters: one-time k-means++ initialization, diagonal covariance matrices (unless otherwise specified), and a maximum of 500 iterations with covariance regularization

of 1×10^{-6}. In practice, the fitting procedure typically converges after approximately 100 iterations. For the selection of the number of modes, we conduct a grid search for ImageNet256 experiments (see Fig. 3) but we also recommend using Bayesian GMM that automatically turns off inactive modes as in Algorithm 1. Specifically, one can first fit a Bayesian GMM on the data with sufficiently large number of modes and then based on the mixture distributions to select the optimal number of modes by discarding the modes with almost-zero weights. For the CLS token GMM (8192 modes with diagonal covariance) that we use for our main results in Tab. 1, it takes approximately 13 hours with peak RAM 289.71GB on an AMD EPYC 7763 64-Core Processor.

10.2. Training Flow Matching Model

As in Tab. 7, we strictly follow DiT$^{\text{DH}}$ [89] for the learning setting, except the core changes that we replace the source distribution and the data coupling. Since we learn a smoother function (see Theorem 3.2), we can afford a larger learning rate of 4.0×10^{-4} and an EMA decay rate of 0.999 for the EMA model to adapt faster. To ensure a fair comparison, we apply the same grid search using same number of epochs (4 epochs) for the Gaussian source baseline. As a result, we use a learning rate 2.0×10^{-4} and an EMA decay rate of 0.995 for Gaussian baseline. However, we come across the same spike loss issues as reported in the official implementation of DiT$^{\text{DH}}$ [89] and LightningDiT [85]. Therefore, after the first 4 epochs, we switch back to the learning rate of 2×10^{-4} and the EMA decay rate of 0.9995 as in Tab. 7 throughout the training.

Table 7. Training configuration of DiT$^{\text{DH}}$-XL with Gaussian and GMM source distribution on DINOv2-B target distribution for Tab. 1.

Component	**Gaussian**	**GMM**
optimizer	AdamW	AdamW
optimizer betas	(0.9, 0.95)	(0.9, 0.95)
optimizer weight decay	0.0	0.0
max learning rate	2×10^{-4}	2×10^{-4}
min learning rate	2×10^{-5}	2×10^{-5}
learning rate decay start epoch	20	20
learning rate decay end epoch	800	800
learning rate decay schedule	linear	linear
batch size	1024	1024
EMA decay rate	0.9995	0.9995
gradient clip	1.0	1.0
loss	$\mathbb{E}_{t,(Z_0,Z_1)\sim\pi} \left\| u_t^\theta\left(Z_t\right) - \left(Z_1 - Z_0\right)\right\|^2$	$\mathbb{E}_{t,(Z_0,Z_1)\sim\pi} \left\| u_t^\theta\left(Z_t\right) - \left(Z_1 - Z_0\right)\right\|^2$
data coupling π	$p(Z_1)p(Z_0)$	$p(Z_1)p(Z_0 \mid Z_1)$
source distribution $p(Z_0)$	$\mathcal{N}(0, I)$	$\sum_{i=1}^m c_i \mathcal{N}(\mu_i, \Sigma_i)$
Diffusion Transformer Backbone	DiT$^{\text{DH}}$-XL	DiT$^{\text{DH}}$-XL
Visual Tokenizer	DINOv2-B	DINOv2-B
Training epochs	80	80
GMM Modes	-	8192 (unless specified otherwise)
Time Distribution Shift	$t_m = \frac{\alpha t_n}{1+(\alpha-1)t_n}$	$t_m = \frac{\alpha t_n}{1+(\alpha-1)t_n}$
Time Distribution Shift α	$\sqrt{\frac{768\times16\times16}{4096}}$	$\sqrt{\frac{768\times16\times16}{4096}}$

Unconditional Generation. The way in which we handle unconditional generation is slightly different from the prior works [51, 62, 89], though might be equivalent. Instead of using a null label by setting the number of classes to 1, we completely eliminate class conditioning (from `c=t+y` to `c=t`) for DiT$^{\text{DH}}$ to learn an unconditional velocity field $u_t^\theta\left(Z_t\right)$.

Conditional Generation. We include a variant of our method with mode conditioning `c=t+y,y=mode` for completeness. During training, we have not applied the condition dropout rate during training. It is feasible to apply it such that during inference time one can use classifier-free guidance (CFG) for mode guidance, but we leave it for future work as we primarily focus on unconditional generation in this paper.

Guidance. We primarily use AutoGuidance [38] as our guidance method because CFG does not apply to the unconditional setting. The idea is to use a weaker diffusion model to guide a stronger one. Following [89], we use the smallest variant, DiT^{DH}-S, training for 20 epochs as the guidance model following the exactly same training configuration in Tab. 7. We consistently use a 1.5 guidance scale for all experiments.

Imbalance Learning. To address the mode imbalance of GMM fitting, we apply focal loss on the objective:

$$\mathcal{L}_{\text{adaptive}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),(z_0,z_1,y)\sim\pi_{\text{mode}}}\left[w(y)\cdot\left\|u_t^\theta(z_t) - (z_1 - z_0)\right\|^2\right] \tag{10.1}$$

where $z_t = (1-t)z_0 + tz_1$, and the adaptive weight $w(y)$ is:

$$w(y) = \text{clamp}\left(\frac{\left(\frac{\bar{L}(y)}{\bar{L}_{\text{mean}}}\right)^\gamma}{\frac{1}{m}\sum_{i=1}^m\left(\frac{\bar{L}(i)}{\bar{L}_{\text{mean}}}\right)^\gamma}, w_{\min}, w_{\max}\right) \tag{10.2}$$

Components:

1. Per-mode loss EMA (tracked over training):

$$\bar{L}(i) \leftarrow \alpha\bar{L}(i) + (1-\alpha)\mathbb{E}_{t\sim\mathcal{U}(0,1),(z_0,z_1)|y=i}\left[\left\|u_t^\theta(z_t) - (z_1 - z_0)\right\|^2\right] \tag{10.3}$$

2. Mean mode loss:

$$\bar{L}_{\text{mean}} = \frac{1}{m}\sum_{i=1}^m \bar{L}(i) \tag{10.4}$$

In our experiments, we used $\gamma = 2.0, w_{\min} = 0.3, w_{\max} = 3.0, \alpha = 0.9995$ and 250 warmup steps (batch size = 1024) to collect statistics before reweighting.

Computation. We use the PyTorch implementation of [89] for all training and inference. With bf16 precision, it takes approximately 21 training hours on an 8×NVIDIA A100-SXM4-40GB node to reach 20 epochs without evaluation.

10.3. Evaluation Metrics

We strictly follow the setup of preprocessing and use the same reference batches of ADM [18] to compare with 50,000 generated images. The FID [33] evaluates the distribution distance of reference and generated images in the Inception-v3 feature space. The IS score [69] utilizes the same Inception-v3 but operates on the raw logit outputs, computing the KL divergence between the overall distribution of predicted labels and the conditional distribution for each image after applying softmax. Precision and Recall [45] are defined in their conventional sense: precision measures what proportion of generated images are realistic, whereas recall measures how much of the training data distribution is captured by the generated samples.

10.4. Data Efficiency Experiments

We construct an ImageNet256 subset with only 10% of the original data by stratified sampling at the class-level. Given the new dataset, we re-train GMM and FM on this subset for both the baseline (Gaussian source distribution with independent coupling) and our method (GMM source distribution with mode coupling) using exactly the same setting as in Tab. 7 as well as the same reference batch by ADM [18] for evaluation.

11. More Experiments

11.1. Linear Probing Results of Various Visual Tokenizers

Qualitatively, the t-SNE visualization (see Fig. 6) shows that all foundation-model-informed visual tokenizers approximately reveal some multimodal structures in their encoded latents compared to standard ones (e.g., SDXL-VAE [63]). Quantitatively, we extract flattened latents from these tokenizers and train a linear classifier preceded by a non-affine BatchNorm1d layer on them for ImageNet-1K training set using SGD with a learning rate of 0.01, momentum of 0.9, no weight decay and a batch size of 1,024 for 50 epochs. The learning rate is halved on plateau (patience of 5 epochs, monitoring top-1 validation accuracy). Both the linear weight and bias are zero-initialized. The reported top-1 and top-5 accuracy on the ImageNet-1K

validation set demonstrate that foundation models as encoders empirically have better probing results than foundation-model-informed visual tokenizers via alignment, distillation and adaptation. We selected DINOv2-B in our experiment due to the availability of corresponding decoders [89] at the time of this project. In Tab. 10, we also show a positive correlation between the linear probing accuracy of visual tokenizers and the gFID of the flow-based generative models employing such visual tokenizers.

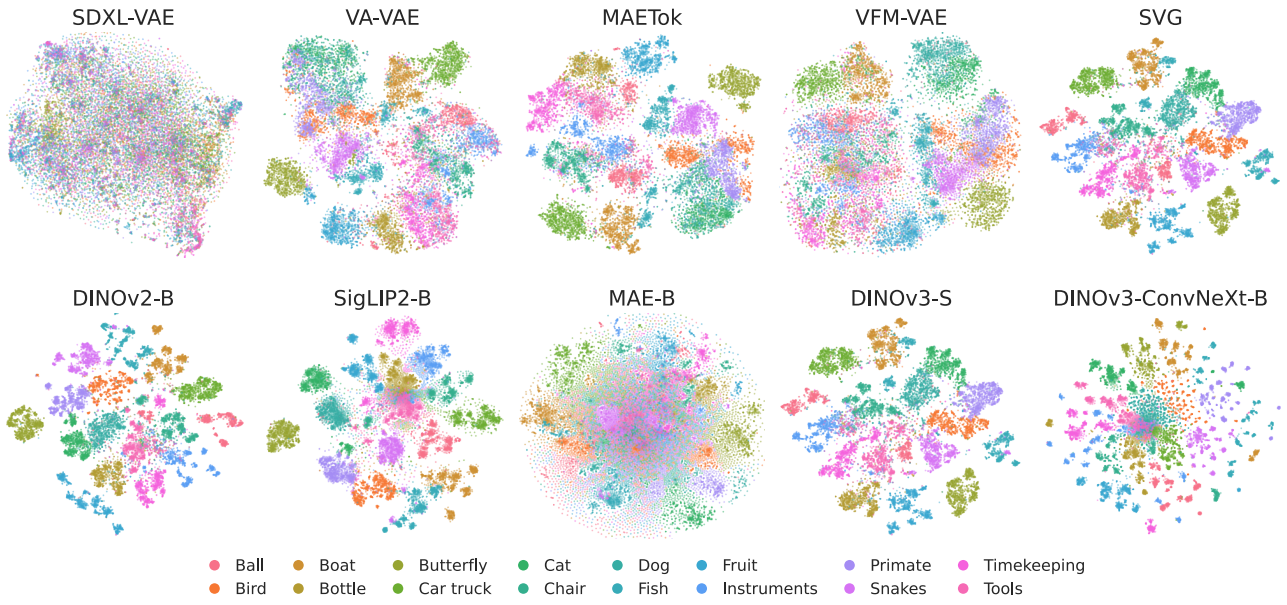

Figure 6. The t-SNE visualization of raw latents colored by 16 superclasses [50]. Importantly, these are raw latent distributions to be learned by the flow-based model, not [CLS] or pooled token representations.

Table 8. The ambient dimension of raw latents and linear probing accuracy on ImageNet-1K validation set for various visual tokenizers. The DINOv2-B's linear probing accuracy is different from Fig. 1 and Tab. 10 as the numbers there are directly taken from [89], but all numbers reported in this table are from the identical learning setup.

	DINOv3-ConvNeXt-B [71]	DINOv2-B [59]	SigLIP2-B [79]	DINOv3-S [71]	SVG-VAE [70]	MAE-B [31]	VFM-VAE [6]	VA-VAE [85]	FLUX.1-VAE [7]	SDXL-VAE [63]
Dim.	65,536	196,608	196,608	98,304	100,352	196,608	8,192	8,192	16,384	4,096
Top-1 (%)	83.29	81.86	80.86	76.70	76.69	64.94	58.17	42.56	7.18	5.02
Top-5 (%)	96.71	96.52	96.26	94.09	94.09	86.20	81.48	69.77	17.06	12.40

11.2. Coarse GMM Estimate is Sufficient

Table 9. The ablative study of covariance and assignment types. We use GMM $m = 64$ for reasonable full covariance estimation, and we train each model for 20 epochs without mode conditioning.

	Soft Assignment			**Hard Assignment**
FID ↓	Spherical	Diagonal	Full	Diagonal
	11.09	**10.44**	10.11	10.67

If we can afford the number of data points, using a more accurate parameterization of covariance captures the local

structures more effectively. However, we observe that the improvement is only marginal (see Tab. 9). Hence, we suggest to prioritize selecting the number of modes (see Fig. 3) over the covariance type. In addition, the soft assignment has performance almost identical to that of the hard assignment. We believe that this is caused by the heavily skewed data-dependent posterior distribution, a phenomenon of concentration of measure in high dimensions.

11.3. Ablation on Foundation Model Encoders

Table 10. The ablative study on foundation model encoders for GMM fitting on ImageNet256 with 8192 modes, diagonal covariance, and soft assignment. We take the linear probing accuracy (%) from [89] and report FID-50K.

Method	Tokenizer	Probing Accuracy↑	Epoch	Prior	Condition	gFID-50K↓
DiT^{DH}-XL	DINOv2-B	84.5%	20	Gaussian	–	16.51
DiT^{DH}-XL + MM-FM	DINOv2-B	84.5%	20	GMM	–	5.11
DiT^{DH}-XL + MM-FM	DINOv2-B	84.5%	20	GMM	Mode	4.74
DiT^{DH}-XL	SigLIP2-B	79.1%	20	Gaussian	–	15.21
DiT^{DH}-XL + MM-FM	SigLIP2-B	79.1%	20	GMM	–	8.27
DiT^{DH}-XL + MM-FM	SigLIP2-B	79.1%	20	GMM	Mode	7.38
DiT^{DH}-XL	MAE-B	68.0%	20	Gaussian	–	27.19
DiT^{DH}-XL + MM-FM	MAE-B	68.0%	20	GMM	–	17.48
DiT^{DH}-XL + MM-FM	MAE-B	68.0%	20	GMM	Mode	16.40

12. Limitations

While our theoretical setup uses some assumptions that may not always hold in practice, we believe these assumptions can be lifted with more careful analysis. For example, while we assume that the prior and target distributions have the same means and covariances, it is easy to allow for mismatch in the parameters, but the inequalities in Theorem 3.1 will have additional terms depending on the parameter error. However, our current analysis gives insight into how the Gaussian mixture prior and mode coupling simplify the generation problem, and we leave refining the analysis to future work.

We primarily focus on unconditional generation in its purest form to intuitively demonstrate our core design insight: flow-based generative models, as data distribution learners, can exploit intrinsic data structure (e.g., multimodality) more efficiently when design choices such as source distribution and data coupling are aligned with that structure. We will leave the extension to conditional generation, such as class condition or text condition that naturally comes with web-scale datasets, to future work, but we would like to provide some thoughts generated during the peer review. In object-centric datasets like ImageNet where foundation models easily get superior zero-shot classification performance, the latent space of these foundation model encoders must be a well-separable multimodal distribution. Having moved to scene-centric web datasets for text-to-image generation (i.e., text-conditioned generation), these foundation model encoders with corresponding trained decoders still shine [78], but the problem is that the target distribution (i.e., latents of these scene-centric data by foundation models) might not be as separable (e.g., underlying modes having shared boundaries) as in object-centric datasets. Therefore, the learning difficulty of the GMM fitting stage becomes non-trivial, not to mention that one has to learn a separate mode predictor $p(\text{mode}|\text{text})$ to supply mode indices during inference. In such a case, by our Theorem 3.1 MM-FM gracefully reduces to classic FM. In other words, MM-FM matches the classic FM in the worst case while offering gains when the structure exists. We suggest the extension shall focus on unified tokenizers that map images and texts to the same space such as SigLIP2 [79]. By doing so, one naturally obtains a mode assignment based on the text latents at inference time, while during training GMM can still be fitted on the image latents, by making the assumption that image latents and text latents of the same semantic meaning are sufficiently close in the unified latent space. For web-scale datasets, fitting a global GMM in one pass may be infeasible, motivating a stochastic mini-batch variant. One natural formulation is bi-level optimization, where the GMM parameters are periodically updated based on downstream FM training signals.

13. Visual Results

(a) Gaussian Source Distribution with Independent Coupling

(b) GMM Source Distribution with Mode Coupling

Figure 7. The qualitative comparison of unconditional generation (50 ODE steps) without guidance between baseline and MM−FM (without mode conditioning) on 10% of ImageNet256 (see Tab. 4) trained for 800 epochs using DiT^{DH}-S. (a) FID=24.33 (b) FID=7.48

(a) Gaussian Source Distribution with Independent Coupling

(b) GMM Source Distribution with Mode Coupling

Figure 8. The qualitative comparison of unconditional generation (50 ODE steps) with guidance between baseline and MM−FM (without mode conditioning) on full ImageNet256 (see Tab. 1) trained for 80 epochs using DiT^{DH}-XL. (a) FID=5.82 (b) FID=3.23