

Learning on the Manifold: Unlocking Standard Diffusion Transformers with Representation Encoders

Amandeep Kumar, Vishal M. Patel

Johns Hopkins University
{akumar99, vpatel136}@jhu.edu
<https://github.com/amandpkr/RJF>

Abstract. Leveraging representation encoders for generative modeling offers a path for efficient, high-fidelity synthesis. However, standard diffusion transformers fail to converge on these representations directly. While recent work attributes this to a capacity bottleneck—proposing computationally expensive “width scaling” of diffusion transformers—we demonstrate that the failure is fundamentally geometric. We identify *Geometric Interference* as the root cause: standard Euclidean flow matching forces probability paths through the low-density interior of the hyperspherical feature space of representation encoders, rather than following the manifold surface. To resolve this, we propose **Riemannian Flow Matching with Jacobi Regularization (RJF)**. By constraining the generative process to the manifold geodesics and correcting for curvature-induced error propagation, RJF enables standard Diffusion Transformer architectures to converge without width scaling. Our method RJF enables the standard DiT-B architecture (131M parameters) to converge effectively, achieving an FID of **3.37** where prior methods fail to converge.

Keywords: Diffusion transformer · Manifold · Geometry

1 Introduction

Flow Matching [1, 7, 18, 19] and Diffusion models [9, 20, 27, 32] have revolutionized generative modeling, enabling high-fidelity synthesis across modalities. While initial approaches operated in pixel space, the paradigm has shifted toward Latent Diffusion Models (LDMs) [27, 33] that leverage compressed representations of VAE [14]. However, because VAEs are optimized for reconstruction, they predominantly capture low-level texture; this forces the diffusion model to learn high-level semantics from scratch, leading to slow convergence. To overcome this, recent works enhance the VAE latent space with semantic priors from strong representation encoders like DiNO [22, 31] and SigLIP [26]. These approaches typically require explicitly aligning semantic representations within the VAE latent space [38] or the diffusion intermediate features [17, 40]. Consequently, these methods require complex auxiliary losses and additional training stages.

Recent work challenges this complexity by proposing Representation Autoencoders (RAE) [42], which discard the VAE entirely in favor of diffusing directly within the feature space of frozen representation encoders. RAEs demonstrate that these high-dimensional semantic representations can support high-fidelity generation without the need for auxiliary alignment losses or complex training stages. However we also have the similar observation like RAE [42] that the standard diffusion recipe fails to converge effectively on these high-dimensional latents, even in a simplified single-image overfitting regime. While RAE attributes this failure to a capacity bottleneck, proposing to scale the transformer width to match the latent dimensionality—we identify a more fundamental cause rooted in the intrinsic geometry of the latent space. We argue that the optimization difficulty arises not from insufficient parameter count, but from a “Geometry Gap”: a structural conflict where the Euclidean probability paths assumed by standard flow matching [18] violate the hyperspherical manifold of representation space.

To understand this failure, we analyze the intrinsic geometry of the feature space. We observe that these high-dimensional representations do not populate the ambient Euclidean space but are strictly confined to a hypersphere, creating a hard shell geometry, and all the information are encoded in angular vectors. We identify the root cause of convergence failure for standard diffusion transformer as **Geometric Interference**: the standard linear probability path used in flow matching cuts through the low-density interior (off-manifold) of hypersphere (forming a chord), rather than following the manifold’s surface [21, 28]. This forces the model to learn a velocity field in regions where the representation space is undefined as shown in Fig. 2. Crucially, we challenge the prevailing hypothesis that this requires scaling model width [42]. Our experiments reveal that the model has sufficient capacity to learn the semantics, but under

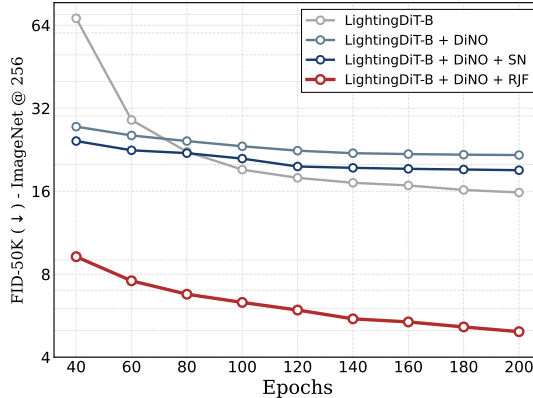


Fig. 1: Bridging the Geometric Gap. We demonstrate that respecting the intrinsic geometry of pre-trained representations encoders enables the use of standard Diffusion Transformers without any architectural modification such as Width Scaling [42]. Our method, Riemannian Flow Matching with Jacobi Regularization (**+DiNO+RJF**), achieves an FID of **4.95** using standard LightingDiT-B [38] architecture without guidance, significantly outperforming the VAE-based LightingDiT-B (FID **15.83**). In contrast, applying standard Flow Matching to DINOv2-B features (**+DiNO**) fails to converge (FID **21.64**) due to *Geometric Interference*. Even restricting the noise to the hypersphere to strictly learn the angular component (**+DiNO+SN**) yields only marginal improvement (FID **19.07**), as Euclidean linear paths still traverse the low-probability interior of the feature manifold.

the standard objective, it wastes its capacity, minimizing radial error (learning feature magnitude which is fixed as the radius of hypersphere) and learning trajectories through the off-manifold area induced by this geometric mismatch.

Motivated from this insight, we propose Riemannian Flow Matching with Jacobi Regularization (RJF). First, we address the trajectory mismatch by adopting Riemannian Flow Matching [4], which replaces the Euclidean linear path with Spherical Linear Interpolation (SLERP). This ensures the generative process follows the geodesic (shortest path along the curve between two points), staying strictly on the manifold surface. Second, we recognize that simply fixing the path is insufficient because the flow matching objective remains geometrically unaware; it treats errors uniformly. On a positively curved hypersphere, velocity errors propagate nonlinearly due to the focusing of geodesics (similar to how parallel longitude lines eventually meet at the poles). To correct this, we introduce a Jacobi Regularization derived from Jacobi fields [41], which reweights the loss to account for curvature-induced distortion. This geometric alignment enables standard DiT architectures to converge efficiently without any architectural modification.

Our contributions are summarized as follows:

- **Geometric Analysis of Convergence Failure:** We identify Geometric Interference as fundamental bottleneck preventing standard diffusion transformers from learning directly on high-dimension representations. We demonstrate that failure arises not only from a capacity deficit, but from the Euclidean objective forcing the model to minimize radial errors and learning trajectories through the low-density interior of the feature manifold.
- **Riemannian Flow Matching with Jacobi Regularization:** We propose a geometrical framework that defines the generative process directly

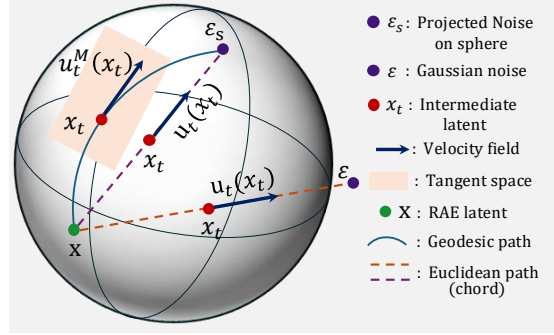


Fig. 2: Geometric Trajectories on the Hypersphere. Visualization of flow matching paths on the high-dimension semantic representation manifold $\mathcal{S}^{d-1}$. Standard Euclidean Flow Matching constructs linear paths that ignore the manifold geometry. Whether targeting standard Gaussian noise ϵ (orange) or projecting noise onto the sphere ϵ_s (purple) to strictly learn the angular component, the linear interpolation forms a chord that cuts through the low-density interior. This forces the model to learn a velocity field in undefined regions regardless of the endpoint. In contrast, Riemannian Flow Matching follows the geodesic (blue curve), ensuring the intermediate state x_t remains strictly on the manifold surface. The resulting velocity field $u_t^M(x_t)$ is correctly defined within the tangent space (pink plane), naturally respecting the geometry of the high-dimensional semantic representations.

on the hyperspherical manifold of the representation encoder. By combining Riemannian Flow Matching (to correct the trajectory) with Jacobi Regularization (to account for geodesic focusing), we ensure the optimization is consistent with both the topology and curvature of the latent space.

- **Efficient Generative Modeling:** We achieve state-of-the-art performance using standard DiT architectures without the need for computationally expensive architectural changes like width scaling. On the 131M-parameter DiT-B, along with RJF and DINOv2-B achieves an FID of **3.37** with guidance and FID of 4.95 without guidance in 200 epochs as shown in Fig. 1, whereas the standard flow matching fails to converge. These gains persist at scale: on DiT-XL, we attain an FID of **3.62** in 80 epochs without guidance, outperforming both the standard flow matching (FID 4.28) and the VAE-based DiT trained with alignment losses (FID 4.29).

2 Geometrical Analysis

Following RAE [42], we investigate the feasibility of directly using pretrained representation encoders within the Diffusion Transformer framework and had similar observation that standard diffusion recipe fail to converge effectively, even in a simplified single-image overfitting. Rather than seeking marginal architectural improvements to address this failure, we aim to answer a more fundamental question: *Why are these high-dimensional, semantically rich representations resistant to the standard Diffusion recipe?* To answer this, we first analyze the intrinsic geometry of the feature space generated by these pretrained representation encoders.

2.1 The Geometry Gap

We analyze the distribution of the final feature vectors $z \in \mathbb{R}^d$ extracted from the DINOv2-B [22] encoder. Decomposing these features into radial and angular components reveals a rigid geometric constraint: the features do not populate the ambient Euclidean space but are explicitly projected onto a hypersphere $\mathcal{S}^{d-1}$ of fixed radius $\sqrt{d}$:

$$z = r \cdot \hat{z}, \quad \text{where } r \approx \sqrt{d} \text{ and } \hat{z} \in \mathcal{S}^{d-1}. \quad (1)$$

As illustrated in Figure 3, the radial component r exhibits near-zero variance due to the ubiquitous application of LayerNorm. This creates a hard shell geometry where all semantic information is encoded exclusively in the angular component $\hat{z}$. This stands in sharp contrast to the standard Gaussian prior used in diffusion models, which assumes a probability mass concentrated in a diffuse shell.

This hyperspherical geometry reveals why the standard flow matching become suboptimal. The standard algorithm constructs a conditional probability path $p_t(x)$ via linear interpolation between the source distribution (Gaussian noise ϵ) and the target data (x):

$$x_t = (1 - t)x + t\epsilon. \quad (2)$$

In Euclidean space, this linear trajectory is optimal. However, on a hyperspherical manifold, this creates a critical distribution shift as also mentioned in [21, 28]. Since ϵ and x are high-dimensional vectors, they are approximately orthogonal ($\epsilon \cdot x \approx 0$). Consequently, the squared norm of the intermediate state x_t follows:

$$\|x_t\|^2 \approx (1-t)^2 \|x\|^2 + t^2 \|\epsilon\|^2. \quad (3)$$

At $t = 0.5$, the norm collapses to $\|x_{0.5}\| \approx \frac{1}{\sqrt{2}}\sqrt{d} \approx 0.7\sqrt{d}$. This implies that the linear flow trajectory x_t does not stay on the manifold $\mathcal{S}^{d-1}$

but rather cuts through the interior of the hypersphere (a chord). This forces the network to learn a velocity field v_t in regions of the feature space that are strictly off manifold for the pretrained representation encoder. The model must essentially hallucinate valid semantic gradients in a region where the representation space is undefined, leading to the convergence failure.

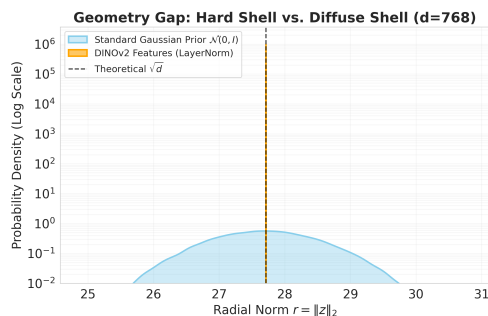


Fig. 3: The Geometry Gap. A comparison of radial feature norms ($r = \|z\|_2$) between DINOv2-B representations and a standard Gaussian prior in $\mathbb{R}^{768}$. While the Gaussian prior (blue) is distributed across a diffuse shell, DINOv2-B features (orange) are rigidly constrained to a hypersphere with near-zero radial variance. This extreme geometric mismatch prevents standard diffusion models from converging effectively on representation encoders.

2.2 Revisiting the Capacity Hypothesis: Geometric Interference

This convergence failure is also identified in RAE [42], to resolve that they proposed a width scaling solution: increasing the Diffusion Transformer’s width (d_{model}) to match at least the token dimension (n). Crucially, they demonstrate that this is not just a capacity issue—simply adding layers (depth) fails to improve convergence. They hypothesize that the bottleneck is strictly dimensional: because the added Gaussian noise is full-rank, a model with width $d_{model} < n$ suffers from rank collapse.

While it is true that a narrow model cannot fully resolve high-dimensional Gaussian noise, but rank collapse should not preclude the learning of the data manifold itself, which often lies on a lower-dimensional subspace [25]. We hypothesized that the failure is not only due to a lack of capacity to model the signal, but rather **Geometric Interference**: the standard Euclidean Flow Matching objective forces the model to prioritize a radial error term that conflicts with representation learning.

To test this, we revisited the single-image overfitting setup. We decomposed flow matching loss into radial (magnitude) and angular (direction) components:

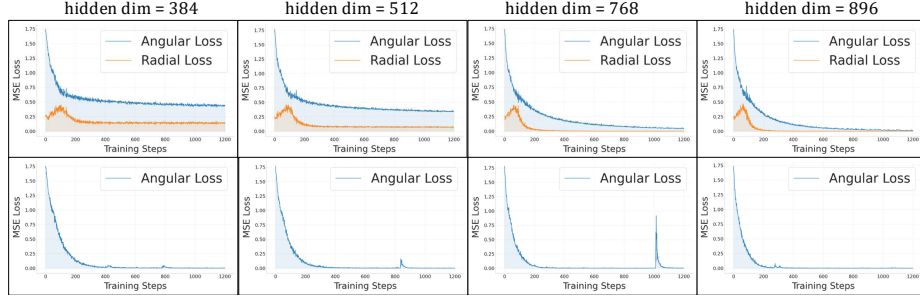


Fig. 4: Geometric Interference vs. Capacity. We train DiT-S models of varying widths on DINOv2 tokens ($d = 768$). **Top Row:** When minimizing Euclidean MSE, narrower models ($d < 768$) suffer from collapse; the Angular Loss (semantics) gets stuck. **Bottom Row:** When the radial loss is ignored, even narrow models ($d = 384$) converge perfectly on the angular component. This proves the bottleneck is not the dimensionality of the data, but the geometric conflict in the objective.

$$\mathcal{L}_{\text{total}} = \underbrace{\|\text{proj}_{\hat{r}}(v_{\text{pred}} - v_{\text{target}})\|^2}_{\text{Radial Loss}} + \underbrace{\|\text{proj}_{\perp}(v_{\text{pred}} - v_{\text{target}})\|^2}_{\text{Angular Loss}}. \quad (4)$$

We then trained DiT models of varying widths ($d = 384$ to $d = 896$) on DINOv2-B tokens ($n = 768$).

As shown in Figure 4, when optimizing the full Euclidean loss (Top Row), models with width $< n$ (e.g., 384, 512) fail completely. The Angular Loss (blue)—which represents the learning of image semantics—stalls and fails to converge. The model effectively wastes its limited rank trying to minimize the Radial Loss (orange), which arises because the Euclidean interpolation forces a chord trajectory that violates the hyperspherical manifold.

However, when we mask the radial loss and optimize *only* the angular component (Bottom Row), the capacity bottleneck vanishes. Even the smallest model ($d = 384$, half the token dimension) converges instantly. This experiment provides a crucial insight: The model has sufficient capacity to learn the semantics, but under Euclidean objective, the radial noise dominates the gradient updates.

The “width scaling” solution proposed by [42] is effectively a brute force fix—it grants the model enough parameters to memorize the ill-posed radial vector field through the void. However, we argue that simply masking the radial component or projecting the noise prior onto the manifold is insufficient to resolve this. While these modifications ensure valid endpoints (isolating the angular component), the underlying Euclidean linear trajectory still forms a chord that traverses the manifold’s interior, as shown in Sec. 2.1. Instead of scaling the architecture to fit a broken objective, we propose to fix the objective itself. By adopting **Riemannian Flow Matching** [4], we define the diffusion process directly on the manifold $\mathcal{S}^{d-1}$. This eliminates the radial conflict by design, en-

sureing that the transport trajectory follows the geodesic (the arc) rather than the chord, naturally aligning the generative process with the representation.

3 Method

3.1 Euclidean Flow Matching

Flow Matching (FM) [18] is a simulation-free framework for training Continuous Normalizing Flows (CNFs). The goal is to learn a time-dependent vector field $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that generates a probability path $p_t(x)$ transforming a simple prior distribution $\epsilon \sim \mathcal{N}(0, I)$ to the complex data distribution $x \sim p_{\text{data}}$.

The flow is defined by the ordinary differential equation:

$$\frac{dx_t}{dt} = v_t(x_t), \quad t \in [0, 1]. \quad (5)$$

To scale this to high dimensions, Conditional Flow Matching (CFM) trains the model to approximate the conditional vector field generating a specific probability path between a data sample $x \sim p_{\text{data}}$ and noise $\epsilon \sim \mathcal{N}(0, I)$.

In the standard Euclidean setting, the simplest probability path is constructed via linear interpolation (Optimal Transport displacement):

$$x_t = (1 - t)x + t\epsilon. \quad (6)$$

Differentiating with respect to time t , the ground-truth conditional velocity field $u_t(x_t|x, \epsilon)$ is constant and straight:

$$u_t(x_t|x, \epsilon) = \frac{d}{dt}((1 - t)x + t\epsilon) = \epsilon - x. \quad (7)$$

The flow matching objective minimizes the mean squared error between the parameterized vector field $v_\theta(x_t, t)$ and the target velocity:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, p(x), p(\epsilon)} [\|v_\theta(x_t, t) - (\epsilon - x)\|^2]. \quad (8)$$

3.2 Riemannian Flow Matching on Hyperspherical Manifolds

While Euclidean Flow Matching has driven recent advances in latent generative modeling, our analysis in Section 2 demonstrates that it is fundamentally ill-suited for the hyperspherical feature spaces produced by representation encoders. The standard linear interpolant violates the manifold structure, forcing the model to learn a vector field through the undefined interior of the sphere.

To resolve this, we propose to reformulate the diffusion process directly on the intrinsic data manifold. We first project our feature vectors to the unit norm hypersphere $\mathcal{M} = \mathcal{S}^{d-1} \subset \mathbb{R}^d$ and define our source distribution by projecting isotropic Gaussian noise onto the manifold $\mathcal{M}$.

Geodesic Probability Paths In the Euclidean setting, the optimal transport path between a source x and target ϵ is a straight line (a chord). On the hypersphere $\mathcal{S}^{d-1}$, the optimal path is the geodesic.

The conditional probability path x_t is defined via Spherical Linear Interpolation (SLERP) rather than linear interpolation. Given data $x \in \mathcal{S}^{d-1}$ and noise $\epsilon \in \mathcal{S}^{d-1}$ (where $\|\epsilon\| = 1$), the geodesic path is given by:

$$x_t = \text{SLERP}(x, \epsilon; t) = \frac{\sin((1-t)\Omega)}{\sin(\Omega)}x + \frac{\sin(t\Omega)}{\sin(\Omega)}\epsilon \quad (9)$$

where $\Omega = \arccos(x^\top \epsilon)$ is the geodesic distance (angle) between the data and the noise. Unlike the Euclidean path, this trajectory ensures that $\|x_t\| = 1$ for all $t \in [0, 1]$, completely eliminating the norm collapse phenomenon and ensuring the generative process on representation manifold.

Tangent Space Velocity Fields A critical consequence of restricting the flow to $\mathcal{M}$ is that the velocity vector v_t must essentially lie in the tangent space $\mathcal{T}_{x_t}\mathcal{M}$ at every point x_t . For the sphere, this implies the velocity must be orthogonal to the position vector: $v_t \cdot x_t = 0$.

The target Riemannian velocity field $u_t^{\mathcal{M}}(x_t|x, \epsilon)$ is computed by differentiating the geodesic path with respect to time t :

$$\begin{aligned} u_t^{\mathcal{M}}(x_t) &= \frac{d}{dt}\text{SLERP}(x, \epsilon; t) \\ &= \frac{\Omega}{\sin(\Omega)} \left(\cos(t\Omega)\epsilon - \cos((1-t)\Omega)x \right). \end{aligned} \quad (10)$$

The Riemannian Objective Consequently, we replace the standard objective with the Riemannian Flow Matching loss. We train the network v_θ to predict this tangent vector field. Crucially, since the target $u_t^{\mathcal{M}}$ lies strictly in the tangent space, the radial component of the error is structurally zero by design. The loss simplifies to the squared norm in the ambient space, which is equivalent to the Riemannian metric induced on the sphere:

$$\mathcal{L}_{\text{RFM}}(\theta) = \mathbb{E}_{t, p(x), p(\epsilon)} [\|v_\theta(x_t, t) - u_t^{\mathcal{M}}(x_t)\|^2]. \quad (11)$$

By optimizing this objective, the model learns purely semantic transitions (angular changes), effectively resolving the Geometric Interference identified in Sec. 2.

To preserve the constant-speed advantage during sampling, we use Geodesic (Exponential Map) Integration. Rather than moving along a straight tangent line that drifts off the manifold, the exponential map wraps the velocity vector around the sphere’s surface.

For a point $x_t \in \mathcal{S}^{d-1}$ and a predicted tangent velocity $v \in \mathcal{T}_{x_t}\mathcal{S}^{d-1}$, the update is defined by the closed-form trigonometric rotation:

$$x_{t+\Delta t} = \cos(\|v\|\Delta t)x_t + \sin(\|v\|\Delta t)\frac{v}{\|v\|}. \quad (12)$$

This update ensures the trajectory follows the great circle exactly, matching the Riemannian flow learned during training. To correct for minor numerical drift over many integration steps, we perform a final rotate and normalize operation. This approach provides a computationally efficient, simulation-free inference path that maintains the rigid DINO geometry without the distortion artifacts of Euclidean solvers.

3.3 Jacobi Field Regularization

While Riemannian Flow Matching with SLERP ensures that the generated path stays on the manifold, the standard velocity-matching objective remains geometrically unaware. The loss $\mathcal{L}_{\text{RFM}} = \|v_\theta - u_t\|^2$ implicitly assumes a flat metric, treating velocity errors uniformly across time $t \in [0, 1]$. However, on a positively curved manifold $\mathcal{S}^{d-1}$, the impact of a velocity error is not uniform. Due to the focusing of geodesics, a perturbation in the velocity vector $w \in \mathcal{T}_{x_t}\mathcal{M}$ propagates non-linearly. To maximize generation fidelity in high dimensional space, we must prioritize minimizing the error near the noise (the endpoint ϵ , $t = 1$).

Inspired by [41], we model this error propagation using Jacobi Fields, which quantify the separation between geodesics caused by velocity perturbations. Solving the Jacobi equation for a hypersphere yields a geometric weighting factor $\lambda(t, \Omega)$ that scales the loss based on the curvature-induced focusing of geodesics:

$$\lambda(t, \Omega) = \text{sinc}^2((1 - t)\Omega), \quad (13)$$

where Ω is the total geodesic distance. This term acts as a geometry-aware attention mechanism: it down-weights errors near $t = 0$ (Data) where geodesic focusing mitigates perturbations, and prioritizes precision near $t = 1$ (noise) where the generative trajectory must precisely align with the feature manifold. The final Jacobi-Regularized objective is:

$$\mathcal{L}_{\text{Jacobi}}(\theta) = \mathbb{E}_{t,x,\epsilon} [\lambda(t, \Omega) \cdot \|v_\theta(x_t, t) - u_t^{\mathcal{M}}(x_t)\|^2]. \quad (14)$$

By optimizing this curvature-corrected objective, we effectively anneal the learning signal, forcing the model to prioritize the learning of high-dimensional latent space. Further details are provided in supplementary.

4 Experiments

4.1 Implementation Details

To ensure a fair comparison, we follow the training protocol of LightingDiT [38]. Experiments are conducted on ImageNet-1K [29] at 256×256 resolution. Unless otherwise specified, we use LightingDiT [38] as our base architecture and train for 80 epochs with a global batch size of 1024. We use the RAE decoder [42] for all representation encoders. Training utilizes the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$) with a fixed learning rate of 2×10^{-4} . We apply gradient clipping with a maximum norm of 1.0 and maintain an Exponential Moving Average (EMA) of weights with a decay of 0.9995. We follow the same setting of dimension dependent noise schedule shift of RAE [42] with $n=4096$. We use Autoguidance [13] as our guidance method with guidance scale of 1.6. For inference, we use an Geodesic (Exponential Map) integrator with 50 steps and evaluate performance on 50k generated images.

4.2 Main Results

Scaling and Training Convergence. We evaluate the convergence and scalability of our method on ImageNet 256×256 generation without guidance, comparing it against DiT [24], LightingDiT [38], REPA [40], and a baseline using DiNOv2-B features with Euclidean Flow Matching (EFM) as shown in Tab. 1. Our method consistently achieves superior FID performance while significantly accelerating convergence across all evaluated model scales. For the DiT-B architecture trained for 80 epochs, our method reduces FID from 21.45 (REPA) and 24.21 (EFM) to 6.77, demonstrating the critical importance of respecting the underlying feature geometry. In the DiT-L setting, we observe a similar trend, where our approach reduces FID from 10.08 to 4.21 compared to

LightningDiT. Notably, in the large-scale setting (DiT-XL), our method demonstrates superior convergence efficiency; at just 24 epochs, it achieves an FID of 6.32, outperforming the strong REPA baseline trained for the full 80 epochs (6.94). By 80 epochs, our method reaches an FID of 3.62, outperforming the Euclidean baseline by 1.19.

State-of-Art Comparison. Due to computational constraints, we benchmark our method in the limited 80-epoch training regime (Table 2). Our LightingDiT -XL model trained with RJF achieves a highly competitive FID of **3.62**, significantly outperforming the Euclidean Flow Matching baseline (FID 4.28) trained on DiNOv2-B features. Crucially, our method demonstrates superior semantic fidelity compared to all other methods. We achieve a state-of-the-art IS of 186.2 and Precision of 0.82, surpassing recent sota methods. This indicates that while our geometric alignment improves FID, it particularly excels at capturing the high-fidelity semantic modes of the data distribution.

In Fig. 6, we present uncured qualitative samples from our LightingDiT-XL model trained with RJF on ImageNet 256×256 . Notably, the model achieves high generation quality and semantic diversity after only 80 epochs of training. More uncured qualitative results are provided in Supplementary.

4.3 Ablation Study

We investigate the impact of each geometric component by training a standard LightingDiT-B/1 model on DiNOv2-B latents. The results are summarized in Table 3. When trained with standard Euclidean Flow Matching (EFM), the

Table 1: FID comparison on ImageNet 256×256 without guidance across various model sizes for LightingDiT with REPA, DiNOv2-B with Euclidean Flow matching (EFM) and RJF.

Model	#Params	Epochs	FID↓
DiT-B/2	130M	80	43.47
LightningDiT-B/1	130M	80	22.86
+ REPA	130M	80	21.45
+ EFM (DiNOv2-B)	131M	80	24.21
+ RJF (DiNOv2-B) (Ours)	131M	80	6.77
DiT-L/2	458M	80	23.33
LightningDiT-L/1	458M	80	10.08
+ REPA	458M	80	7.48
+ EFM (DiNOv2-B)	459M	80	6.31
+ RJF (DiNOv2-B) (Ours)	459M	80	4.21
DiT-XL/2	675M	80	19.47
LightningDiT-XL/1	675M	80	9.29
+ REPA	675M	80	6.94
+ EFM (DiNOv2-B)	677M	14	10.23
+ EFM (DiNOv2-B)	677M	24	7.93
+ EFM (DiNOv2-B)	677M	80	4.28
+ RJF (DiNOv2-B) (Ours)	677M	14	8.83
+ RJF (DiNOv2-B) (Ours)	677M	24	6.32
+ RJF (DiNOv2-B) (Ours)	677M	80	3.62

Table 2: Class-conditional performance on ImageNet 256×256 with and without guidance. Our method achieves a superior FID of **3.62**, outperforming the standard flow matching baseline (FID 4.28).

Method	Epochs	#Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			FID↓	IS↑	Prec.↑	Rec.↑	FID↓	IS↑	Prec.↑	Rec.↑
<i>Pixel Diffusion</i>										
ADM [6]	400	554M	10.94	101.0	0.69	0.63	3.94	215.8	0.83	0.53
RIN [11]	480	410M	3.42	182.0	-	-	-	-	-	-
PixelFlow [5]	320	677M	-	-	-	-	1.98	282.1	0.81	0.60
PixNerd [34]	160	700M	-	-	-	-	2.15	297.0	0.79	0.59
SiD2 [10]	1280	-	-	-	-	-	1.38	-	-	-
<i>Vanilla Latent Diffusion</i>										
DiT [24]	1400	675M	9.62	121.5	0.67	0.67	2.27	278.2	0.83	0.57
MaskDiT [43]	1600	675M	5.69	177.9	0.74	0.60	2.28	276.6	0.80	0.61
SiT [20]	1400	675M	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
TREAD [16]	740	675M	-	-	-	-	1.69	292.7	0.81	0.63
MDTV2 [8]	1080	675M	-	-	-	-	1.58	314.7	0.79	0.65
<i>Latent Diffusion with Self-supervised Representation Model</i>										
REPA [40]	800	675M	5.90	157.8	0.70	0.69	4.70	305.7	0.80	0.65
REPA-E [17]	800	675M	1.83	217.3	-	-	1.26	314.9	0.79	0.66
REG [37]	480	677M	2.20	219.1	0.77	0.66	1.40	296.9	0.77	0.66
LightningDiT [38]	800	675M	2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
DDT [35]	800	675M	6.27	154.7	0.68	0.69	1.26	310.6	0.79	0.65
RAE (DiT ^{OH}) [42]	800	839M	1.60	242.7	0.79	0.65	1.28	262.9	0.78	0.67
SFD (XL) [23]	800	676M	2.54	-	-	-	1.06	267.0	0.78	0.67
REPA [40]	80	675M	7.90	122.6	0.70	0.65	-	-	-	-
REPA-E [17]	80	675M	3.46	159.8	0.77	0.63	1.67	266.3	0.80	0.63
REG [37]	80	677M	3.40	184.1	0.77	0.63	1.86	321.4	0.76	0.63
LightningDiT [38]	64	675M	5.14	130.2	0.76	0.62	2.11	252.3	0.81	0.58
SVG [30]	80	675M	6.57	137.9	-	-	3.54	207.6	-	-
SFD (XL) [23]	80	675M	3.53	162.0	0.75	0.65	1.30	233.4	0.78	0.64
DiT-XL(DiNOv2-B) [38]	80	677M	4.28	-	-	-	-	-	-	-
DiT-XL(DiNOv2-B) + RJF (Ours)	80	677M	3.62	186.2	0.82	0.52	2.81	201.22	0.82	0.56

model fails to converge effectively, yielding a poor FID of **24.32**. As analyzed in Sec. 2, baseline suffers from Geometric Interference: the model wastes capacity minimizing radial errors and learning velocity fields within the undefined interior of the manifold.

To address the radial error, we project both the Gaussian noise and the target latents onto the unit sphere, effectively forcing the model to learn only the angular component. We observe only a marginal improvement of 2.33 FID (reaching **21.99**). This indicates that simply fixing the endpoints is insufficient; the issue persists because the Euclidean linear interpolation still forms a chord that violates the manifold geometry.

By adopting Riemannian Flow Matching to ensure the generative trajectory follows the geodesic, we observe a huge performance improvement. The FID drops significantly to **7.06** (with IS improving to 136.70). This result is critical: it validates our claim that respecting intrinsic geometry enables standard Diffusion Transformers to model high-dimensional features

Table 3: Ablation of Geometric Components. We train a LightningDiT-B model on DiNOv2-B features. The Standard Euclidean baseline fails to converge (FID 24.32) due to geometric interference. Projecting noise to sphere (+SN) yields only marginal gains, as the linear path remains flawed. Adopting **Riemannian Flow Matching** (+RFM) resolves the trajectory mismatch, drastically improving FID to 7.06, with **Jacobi Regularization** (+RJF), achieves SOTA performance (FID **6.77**), demonstrating that respecting geometry eliminates need for width scaling, further training leads to FID **3.37**.

Method	Epochs	FID↓	IS↑	Prec.↑	Rec.↑
DiT-B/1(DiNOv2-B)	80	24.32	79.34	0.63	0.46
DiT-B/1(DiNOv2-B) + SN	80	21.99	98.25	0.62	0.47
DiT-B/1(DiNOv2-B) + RFM	80	7.06	136.70	0.78	0.49
DiT-B/1(DiNOv2-B) + RJF	80	6.77	138.12	0.78	0.50
DiT-B/1(DiNOv2-B) + RJF	200	4.95	157.48	0.79	0.52
DiT-B/1(DiNOv2-B) + RJF w/ guid	200	3.37	180.26	0.80	0.56

without the need for architectural width scaling. Finally, incorporating Jacobi Regularization to weight the loss according to geodesic focusing further refines the geometric alignment, improving FID to **6.77**. When trained for 200 epochs, our method (RJF) achieves a remarkable FID of **4.95**. With guidance, the performance reaches SOTA with an FID of **3.37** and an IS of **180.26**.

5 Discussion

5.1 Generality Across Architectures

To validate the robustness of our approach, we evaluate RJF across diverse Diffusion Transformer architectures (Table 4).

We observe consistent performance gains in all settings. On the large-scale LightingDiT-XL, our method achieves an FID of **3.62**, significantly outperforming both the standard VAE-based baseline (FID 4.29) and the Euclidean Flow Matching on DINOv2 features (FID 4.28). This confirms that our gains scale effectively to larger models. We observe similar improvements on the DDT-XL architecture, where RJF lowers the FID to **5.82**, surpassing the Euclidean baseline (FID 6.55). We also evaluate $DiT^{DH} - S$, an architecture explicitly designed with width scaling to handle RAE latents. Even in this specialized setting, our geometric objective yields further improvements (reaching FID **6.20**).

Table 4: Performance with different architecture design with DiNOV2-B

Method	FID↓
DiT-XL/1	4.29
DiT-XL/1	4.28
DiT-XL/1 + RJF (ours)	3.62
DDT-XL/2	6.62
DDT-XL/1	6.55
DDT-XL/1 + RJF (ours)	5.82
DiT^{DH}	6.33
DiT^{DH} + RJF (ours)	6.20

5.2 Different representation encoders

To assess the generalization of our method beyond DINOv2, we evaluate RJF on two distinct representation encoders: SigLIP (contrastive) and MAE (reconstructive). As shown in Table 5, the standard LightingDiT-B/1 baseline fails to converge effectively on either representation, yielding poor FIDs of **130.21** for SigLIP and **50.48** for MAE.

We attribute this failure to shared geometric properties of these latent spaces. SigLIP is trained with a contrastive objective that explicitly enforces hyperspherical normalization [36]. Similarly, while MAE is reconstructive, the extensive application of LayerNorm strictly constrains its features to a hyperspherical manifold. Consequently, our method respects intrinsic geometry resolves the convergence issues in both cases. RJF achieves a dramatic improvement, reaching an FID of **10.39** on SigLIP and **19.82** on MAE, demonstrating that geometric alignment is essential for learning on representation feature spaces.

Table 5: Ablation study with different Representation encoder

Method	FID (↓)	
	SIGLIP	MAE
DiT-B/1	130.21	50.48
DiT-B/1 + RJF	10.39	19.82

5.3 Projection with different radius

We further analyze the impact of the projection radius R during the inference stage (Fig. 5). While the model is trained on the intrinsic geometry, we observe that strictly reprojecting the generated latents back to the original DINOv2-B norm ($R \approx 27.7$) yields a suboptimal FID of 7.79. Interestingly, increasing the projection radius during inference leads to consistent performance gains, achieving the best FID of **6.77** at a radius of $R \approx 45$. This observation suggests that the RAE decoder is sensitive to feature magnitude; amplifying the norm of the generated latents leading to enhance the fidelity of the generated images.

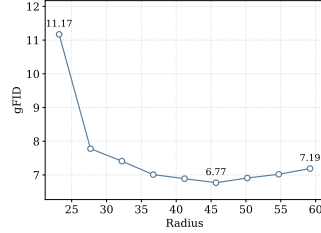


Fig. 5: Performance across different radius projection

6 Related Works

6.1 Representation Alignment for image generation

Recent research has increasingly focused on bridging the gap between generative models and pretrained representation encoders to enhance diffusion transformer performance. A foundational approach in this domain involves feature-space alignment, where methods like REPA [40] accelerate convergence by aligning intermediate diffusion features with pretrained representations such as DINOv2 [22]. This paradigm has been extended by architectures like DDT [35], which applies alignment to a decoupled encoder-decoder structure, and REG [37], which introduces a learnable class token for explicit semantic guidance. Beyond feature alignment, significant efforts have been made to optimize the latent space itself; REPA-E [17] enables joint optimization of the VAE and diffusion model, ReDi [15] jointly learns low-level and high-level semantic distributions, and approaches like VA-VAE [38], and SVG [30] enrich the conventional VAE with pretrained visual encoder representations. Most recently,

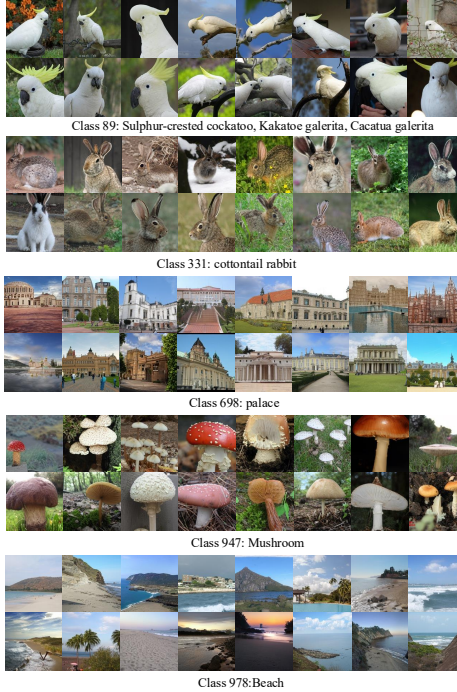


Fig. 6: Qualitative results of LightingDiT-XL+RJF trained for 80 epochs on ImageNet 256×256 . We show uncurated results on the five classes.

Most recently,

RAE [42] proposes replacing the VAE entirely with representation encoders. However, to mitigate the training convergence failure associated with this approach, RAE relies on width scaling. While we similarly observe that standard diffusion recipes fail in this setting, we attribute this not due to model capacity, but to a fundamental geometric mismatch. We demonstrate that standard Euclidean flow matching trajectories inadvertently traverse the low-density interior of the feature hypersphere, necessitating a geometrically consistent modeling approach rather than incremental architectural modifications.

6.2 Riemannian Flow Matching.

While diffusion models rely on stochastic differential equations, Flow Matching (FM) [18] has emerged as a robust, simulation-free alternative for training Continuous Normalizing Flows (CNFs). Standard FM constructs probability paths via linear interpolation in Euclidean space, regressing a velocity field to guide samples from a source distribution to the data. However, as noted in recent geometric deep learning literature [21, 28], Euclidean linear paths are ill-suited for data residing on non-Euclidean manifolds, as they violate the intrinsic geometry of the domain. To address this, Riemannian Flow Matching (RFM) [4] generalizes the framework by replacing Euclidean straight lines with geodesic paths defined by the Riemannian metric. By defining the conditional probability path as a geodesic interpolation, RFM ensures that the flow remains strictly on the manifold, avoiding regions of low density such as the interior of a hypersphere. This formulation has been widely adopted for generative modeling on structured scientific domains, including protein backbone generation on $SE(3)$ [2, 39], torsion angle prediction on tori [12], and motion planning for robotics on configuration manifolds [3].

7 Conclusion

In this work, we demonstrate that the failure of standard diffusion transformer on representations encoder (like DiNO, SigLip and MAE) is not a capacity issue, but geometric. We identified that standard Euclidean objectives suffer from Geometric Interference, wasting computation on the minimizing radial error and learning trajectories through the off-manifold area induced by this geometric mismatch. We solved this with Riemannian Flow Matching with Jacobi Regularization (RJF), which enforces geodesic trajectories consistent with the latent topology. Crucially, RJF unlocks the standard DiT-B architecture (131M parameters), achieving a state-of-the-art FID of 3.37 where baselines fail to converge. We further showed that this geometric alignment scales efficiently to DiT-XL (FID 3.62) in just 80 epochs, establishing that efficient generation requires respecting the latent topology rather than just scaling model width.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2022)
2. Bose, A.J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C.H., Nica, A.C., Korablyov, M., Bronstein, M., Tong, A.: Se (3)-stochastic flow matching for protein backbone generation. arXiv preprint arXiv:2310.02391 (2023)
3. Braun, M., Jaquier, N., Rozo, L., Asfour, T.: Riemannian flow matching policy for robot motion learning. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5144–5151. IEEE (2024)
4. Chen, R.T., Lipman, Y.: Flow matching on general geometries. arXiv preprint arXiv:2302.03660 (2023)
5. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
8. Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Mdtv2: Masked diffusion transformer is a strong image synthesizer. arXiv preprint arXiv:2303.14389 (2023)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
10. Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., Salimans, T.: Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. arXiv preprint arXiv:2410.19324 (2024)
11. Jabri, A., Fleet, D., Chen, T.: Scalable adaptive computation for iterative generation. arXiv preprint arXiv:2212.11972 (2022)
12. Jing, B., Corso, G., Chang, J., Barzilay, R., Jaakkola, T.: Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems* **35**, 24240–24253 (2022)
13. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems* **37**, 52996–53021 (2024)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. arXiv preprint arXiv:2504.16064 (2025)
16. Krause, F., Phan, T., Gui, M., Baumann, S.A., Hu, V.T., Ommer, B.: Tread: Token routing for efficient architecture-agnostic diffusion training. arXiv preprint arXiv:2501.04765 (2025)
17. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
19. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)

20. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
21. Mathieu, E., Nickel, M.: Riemannian continuous normalizing flows. *Advances in neural information processing systems* **33**, 2503–2515 (2020)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
23. Pan, Y., Feng, R., Dai, Q., Wang, Y., Lin, W., Guo, M., Luo, C., Zheng, N.: Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion. *arXiv preprint arXiv:2512.04926* (2025)
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
25. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894* (2021)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Rozen, N., Grover, A., Nickel, M., Lipman, Y.: Moser flow: Divergence-based generative modeling on manifolds. *Advances in neural information processing systems* **34**, 17669–17680 (2021)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**(3), 211–252 (2015)
30. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301* (2025)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
32. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020)
33. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. *Advances in neural information processing systems* **34**, 11287–11302 (2021)
34. Wang, S., Gao, Z., Zhu, C., Huang, W., Wang, L.: Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268* (2025)
35. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741* (2025)
36. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
37. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467* (2025)

- 38. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025)
- 39. Yim, J., Campbell, A., Foong, A.Y., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V.G., Veeling, B.S., Barzilay, R., Jaakkola, T., et al.: Fast protein backbone generation with se (3) flow matching. arXiv preprint arXiv:2310.05297 (2023)
- 40. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
- 41. Zaghen, O., Eijkelboom, F., Pouplin, A., Liu, C., Welling, M., van de Meent, J.W., Bekkers, E.J.: Riemannian variational flow matching for material and protein design. arXiv preprint arXiv:2502.12981 (2025)
- 42. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
- 43. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers. arXiv preprint arXiv:2306.09305 (2023)