The Geometry of Generative Latents: Navigability and Invertibility

Zexi Jia[1] **Jingwei Li**[2] **Pengcheng Luo**[2] **Shipeng Cen**[2]
Longhui Yuan[1] **Jinchao Zhang**[1] **Jie Zhou**[1]
[1]WeChat AI, Tencent Inc, China
[2]Institute for Artificial Intelligence, Peking University

Abstract

Latent diffusion models depend critically on the geometry of the space in which generation is performed. A persistent challenge in VAE- and RAE-based systems is that detail-rich latents improve inversion but can make diffusion dynamics harder to model. We study this reconstruction–generation tension through two geometric properties: *invertibility*, which measures how faithfully images can be decoded, and *navigability*, which measures how smoothly diffusion can transport mass in latent space. We introduce a tangent-projected transport complexity that isolates score-field variation along data-manifold directions and highlights intrinsic dimension, curvature, reach, and enrichment direction as geometry-side factors in diffusion difficulty. This analysis yields the Normal Projection Principle: when enriching a navigable base latent, added detail should lie in the normal space so that the base coordinate is preserved under tubular projection, whereas tangential residuals shift that coordinate at first order. Motivated by this principle, we introduce NormFuse, an RAE-based text-to-image model that fuses frozen multi-layer DINOv3 features through normal-space projection. Controlled and large-scale matched comparisons show that NormFuse preserves reconstruction benefits while avoiding the generation degradation of unconstrained fusion, supporting latent geometry as both an explanatory lens and a practical design rule.

1 Introduction

Latent diffusion models generate images by learning the denoising process in a representation space rather than directly in pixels [Peebles and Xie, 2023, Esser et al., 2024]. The latent space is therefore not merely a compression module in the pipeline; it is one of the central factors that determines both the fidelity of the decoded image and the difficulty of the generative modeling problem itself.

This has become increasingly visible in recent work. Across VAEs, visual tokenizers, and frozen vision foundation model (VFM) features used in RAE-style pipelines, improving reconstruction does not necessarily improve generation [Yao and Wang, 2025, Kouzelis et al., 2025, Skorokhodov et al., 2025, Zheng et al., 2025, Shi et al., 2025]: representations that preserve more local detail often yield better reconstructions but lead to worse sample quality or slower optimization. These observations point to a deeper question: what actually makes a latent space good for generation?

In this paper, we use geometry as a lens for this question. A useful generative latent must satisfy two competing requirements. First, it should be *invertible*: a decoder should be able to recover image content from the latent with low error. Second, it should be *navigable*: diffusion should be able to learn smooth transport over the latent distribution. These two requirements favor different latent geometries. Invertibility often benefits from richer representations that preserve fine-grained structure, while navigability favors spaces that are smoother, better connected, and easier for diffusion to traverse. From this perspective, the familiar reconstruction–generation dilemma reflects a structural

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

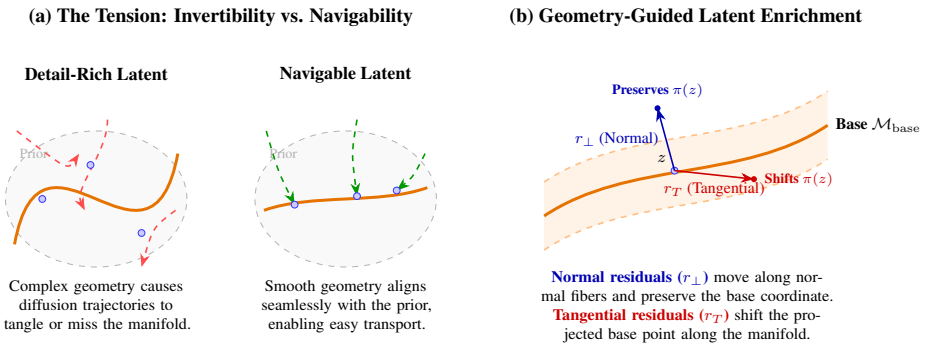


Figure 1: **A geometric view of generative latent quality. (a) The Reconstruction–Generation Dilemma.** Latents must balance *invertibility* (recovering detail) and *navigability* (smooth diffusion transport). Maximizing invertibility (left) often forms highly curved manifolds that tangle diffusion trajectories, while highly navigable latents (right) align smoothly with the prior but discard detail. **(b) Geometry-Guided Enrichment.** When adding detail to a base latent, normal residuals ($r_\perp$) move along normal fibers and preserve the base coordinate under tubular projection. Tangential residuals (r_T), in contrast, shift the projected base point along the manifold at first order, mixing added detail with the transport directions of the base latent.

tension in the geometry of generative representations, and improving a latent requires attention not only to how much detail is added but also to the direction along which it is added. We instantiate this view in an RAE-based text-to-image model built on frozen DINOv3 features that preserves the base latent coordinate while recovering complementary detail.

Our contributions are as follows:

- **We propose a geometric view of generative latent quality.** We frame a good generative latent as one that must balance *invertibility* and *navigability*, providing a unified perspective on the reconstruction–generation trade-off observed across latent diffusion systems.
- **We derive a geometry-guided principle for latent enrichment.** We analyze how latent manifold geometry, characterized by intrinsic dimension, curvature, and reach, and the direction of detail injection affect the geometry-side complexity faced by diffusion. This yields the *Normal Projection Principle*, namely that added detail should lie in the normal space of the base manifold, which we sanity-check on synthetic manifolds with known geometry.
- **We build an RAE-based text-to-image model, NormFuse.** Using frozen DINOv3 features, we show under controlled ablations that normal-projected fusion avoids the degradation caused by unconstrained fusion while preserving reconstruction quality. We further add matched large-scale RAE-Base and Predictor-Add comparisons, trained with the same data and training recipe, and show that NormFuse improves compositional text-to-image alignment over both the single-layer RAE baseline and unconstrained multi-layer residual fusion.

2 Related Work

Latent reconstruction and representation autoencoders. A recurring finding in latent diffusion is that better reconstruction does not necessarily translate into better generation. This tension has been observed across VAEs, visual tokenizers, and representation-based pipelines: increasing latent capacity or preserving more local detail can improve reconstruction while degrading sample quality or making diffusion harder to optimize [Yao and Wang, 2025, Kouzelis et al., 2025, Skorokhodov et al., 2025]. RAE-style methods replace the learned VAE encoder with frozen vision foundation model features [Zheng et al., 2025], and SVG-T2I scales this idea to text-to-image generation with DINOv3 features and an unconstrained residual branch for detail recovery [Shi et al., 2025]. Existing variants mainly alter encoder layer, decoder capacity, or residual predictor, but do not explicitly control whether added detail moves along or away from the base representation manifold. Related methods such as REPA and Latent Forcing align or separate representations at different stages of the generative stack [Yu et al., 2025, Baade et al., 2026].

Latent geometry and diffusion difficulty. The manifold view of high-dimensional data has long motivated representation learning and generative modeling [Pope et al., 2021]. Recent theory for score-based diffusion relates sampling and score estimation to intrinsic dimension, curvature, and reach [De Bortoli et al., 2022, De Bortoli, 2022], while empirical work emphasizes interpolation, local connectivity, and representation geometry in realistic generation [Kamb and Ganguli, 2025, Somepalli et al., 2023, Aithal et al., 2024, Xu et al., 2026]. Classical manifold-learning methods estimate local tangent spaces for dimension reduction or representation analysis; in contrast, we use local PCA to turn a tangent–normal decomposition into an architectural constraint for multi-layer VFM fusion. This gives a concrete design rule: residual detail should be projected onto the normal space of a navigable base token manifold.

3 Geometric Framework

3.1 Two Competing Requirements: Navigability and Invertibility

Consider a latent diffusion pipeline in which an encoder $E : \mathcal{X} \to \mathcal{Z}$ maps images to a latent space $\mathcal{Z} \subset \mathbb{R}^D$, a decoder $G : \mathcal{Z} \to \mathcal{X}$ reconstructs images from latents, and a diffusion model is trained to generate samples from the induced latent distribution $p(z)$. For this pipeline to produce both diverse and faithful images, the latent space must satisfy two requirements that are in fundamental geometric tension. For spatial RAE latents, we apply this framework to token vectors: each image latent is a grid of token-level coordinates, and the manifold model describes the marginal distribution of these token vectors in $\mathbb{R}^D$. The full image latent lies on a spatially constrained submanifold of the product of token manifolds, so token-wise normal projection should be read as a local product-space approximation rather than a complete model of all spatial correlations.

Navigability. For the theoretical diagnostic, write $s_t(z) = \nabla_z \log p_t(z)$ for the population score at noise level t. Smooth score variation motivates a proxy for approximation difficulty under a fixed model, rather than a necessary-and-sufficient criterion for tractable generation. We summarize navigability through the integrated score curvature:

$$\mathrm{Nav}(\mathcal{Z}) \;:=\; \int_{t_{\min}}^{T} \mathop{\mathbb{E}}_{z\sim p_t} \Big[\big\| \nabla_z s_t(z) \big\|_F^2 \Big]\, dt. \tag{1}$$

In practice, diffusion models are trained on a finite effective noise range, and we use $t \in [t_{\min}, T]$ throughout the analysis. Smaller values indicate smoother score fields and serve here as a geometry-side proxy for easier diffusion training under fixed model and optimization conditions. Although our experiments use flow matching rather than explicit score matching, its learned velocity field is distinct from this population score. We use score-field curvature only as a qualitative proxy for target-field roughness and finite-model approximation difficulty, not as a direct optimization bound.

Invertibility. The decoder must recover image content with small error:

$$\mathrm{Inv}(\mathcal{Z}) \;:=\; \mathop{\mathbb{E}}_{x\sim p_{\mathrm{data}}} \big[\ell\big(x, G(E(x))\big) \big], \tag{2}$$

where ℓ is a reconstruction loss. Thus Inv is an inversion error, and lower values indicate better invertibility. Retaining fine detail can increase effective latent complexity, although low reconstruction error does not mathematically require a higher intrinsic dimension.

The fundamental tension. Navigability favors a low-dimensional, gently curved, well-connected manifold. Invertibility favors higher intrinsic dimension and more complex local geometry. This tension surfaces as the reconstruction–generation dilemma in VAEs [Yao and Wang, 2025], the layer trade-off in VFMs [Zheng et al., 2025, Shi et al., 2025], and evidence that interpolated FID predicts generation quality more reliably than reconstruction FID alone [Xu et al., 2026]. The question is whether invertibility can be improved without paying the full cost in navigability.

3.2 Tangent-Projected Transport Complexity: Linking Geometry to Diffusion Difficulty

Since $s_t(z) = \nabla_z \log p_t(z)$, the score Jacobian in Eq. 1 coincides with the Hessian of the log-density, $\nabla_z^2 \log p_t(z)$. A raw ambient Hessian, however, also contains universal normal curvature introduced

by the Gaussian forward process. Even for a flat low-dimensional manifold embedded in $\mathbb{R}^D$, smoothing creates a normal quadratic log-density whose Hessian scales with the codimension. This component is not specific to the geometry that diffusion must navigate along the data support. We therefore measure the score curvature along the local tangent directions of the latent manifold.

Let $\mathcal{M} \subset \mathbb{R}^D$ be a smooth latent manifold with reach $\tau > 0$. For z in the tubular neighborhood of $\mathcal{M}$, let $\pi(z)$ be the nearest point on $\mathcal{M}$ and define $\Pi_T(z) = \Pi_{T_{\pi(z)}\mathcal{M}}$. We analyze the finite training-noise range $t \in [t_{\min}, T]$, where $t_{\min} > 0$ is the smallest effective noise level used by the diffusion objective. The projector only needs to be defined inside the tubular neighborhood; values outside the tube are excluded by the indicator and absorbed into the tail term.

Definition 3.1 (Tangent-projected transport complexity)**.** The *tangent-projected transport complexity* of a latent space $\mathcal{Z}$ with data distribution $p(z)$ is

$$\mathcal{C}_T(\mathcal{Z}) \;=\; \int_{t_{\min}}^{T} \mathop{\mathbb{E}}_{z\sim p_t}\Big[\big\|\Pi_T(z)\,\nabla_z^2 \log p_t(z)\,\Pi_T(z)\big\|_F^2\,\mathbb{1}\{z \in \mathcal{T}_\tau(\mathcal{M})\}\Big]\,dt \;+\; R_{\text{tail}}, \tag{3}$$

where p_t denotes the marginal at noise level t under the standard Gaussian forward process. The indicator restricts the projected Hessian to the tubular neighborhood where π and Π_T are well defined, and R_{tail} absorbs the contribution outside this neighborhood.

Lower $\mathcal{C}_T$ indicates smoother score variation along the directions in which data mass is transported, and we use it as a geometry-side proxy for easier diffusion under fixed model, optimization, data, and sampling conditions. It is not a standalone guarantee of better sample quality. The following proposition relates this proxy to geometric properties of the latent manifold.

Proposition 3.2 (Geometry bounds tangent-projected transport complexity)**.** *Suppose the support of $p(z)$ concentrates on a smooth compact C^3 submanifold $\mathcal{M} \subset \mathbb{R}^D$ of intrinsic dimension d, with sectional curvature bounded by $|\kappa_{\text{sec}}| \le K$ and reach $\tau > 0$. Assume that the density with respect to the Riemannian volume measure is bounded above and below by positive constants, with uniformly bounded first two derivatives. Then, for any fixed $0 < t_{\min} < T$,*

$$\mathcal{C}_T(\mathcal{Z}) - R_{\text{tail}} \;\le\; C \cdot d^2 \cdot \Big(1 + \frac{K}{\tau^2} + \frac{1}{\tau^4}\Big), \tag{4}$$

where $C > 0$ absorbs the density bounds, derivative bounds, the volume of $\mathcal{M}$, the finite noise interval, and the constants needed to express all quantities in normalized latent units. The displayed dependence should be read as the bound on the tubular contribution $\mathcal{C}_T - R_{\text{tail}}$; when the tail is controlled it can be absorbed into C. It removes the explicit ambient-Hessian codimension factor, while any residual contribution from samples outside the tubular neighborhood is contained in R_{tail}.

The bound is qualitative, but its structural message is useful: after tangent projection, the flat codimension-dominated Gaussian term is removed, while curvature-induced tangent terms remain controlled by intrinsic dimension d, curvature K, and reach τ rather than by ambient dimension alone. All quantities are expressed in the normalized latent units used for diffusion training; for manifold families whose curvature contractions or mean curvature grow with d, additional dimension-dependent constants can enter C. Thus the reconstruction–generation dilemma can arise when improving reconstruction increases intrinsic dimension or decreases reach. We use this result as a geometry-side diagnostic and design motivation, not as a direct guarantee on FID or on the flow-matching training loss; further scope discussion and proof details are given in Appendix B. Proposition 3.2 answers a global question about what makes a latent navigable. We next ask a local design question: given a navigable base latent, how can we add detail while minimally disturbing its coordinate system?

3.3 Directional Analysis and the Normal Projection Principle

Proposition 3.2 suggests that manifold geometry is one important contributor to diffusion difficulty. We now ask: when adding detail to a navigable base latent, does the *direction* of enrichment matter?

Given a base latent $u \in \mathbb{R}^D$ on $\mathcal{M}_{\text{base}}$ and a detail residual $r \in \mathbb{R}^D$, decompose the residual into tangent and normal components:

$$r = r_T + r_\perp, \qquad r_T = \Pi_{T_u\mathcal{M}_{\text{base}}}(r), \qquad r_\perp = \Pi_{N_u\mathcal{M}_{\text{base}}}(r). \tag{5}$$

Let $\tau_0 > 0$ denote the reach of the base manifold $\mathcal{M}_{\text{base}}$, and let $\mathcal{T}_{\tau_0}(\mathcal{M}_{\text{base}}) = \{z \in \mathbb{R}^D : \text{dist}(z, \mathcal{M}_{\text{base}}) < \tau_0\}$ denote its tubular neighborhood. We formalize the effect of each component through the nearest-point projection $\pi : \mathcal{T}_{\tau_0}(\mathcal{M}_{\text{base}}) \to \mathcal{M}_{\text{base}}$, which is well defined and smooth on $\mathcal{T}_{\tau_0}(\mathcal{M}_{\text{base}})$. For a fusion weight $\alpha > 0$, define the augmented latent

$$z_\alpha(u) = u + \alpha\, r(u), \qquad \alpha \sup_{v \in \mathcal{M}_{\text{base}}} \|r(v)\|_2 < \tau_0/2. \tag{6}$$

Proposition 3.3 (Directional effect under tubular projection)**.** *Let $\mathcal{M}_{\text{base}} \subset \mathbb{R}^D$ be a smooth compact submanifold with reach $\tau_0 > 0$, and let $\pi : \mathcal{T}_{\tau_0}(\mathcal{M}_{\text{base}}) \to \mathcal{M}_{\text{base}}$ be the nearest-point projection. For $z_\alpha(u) = u + \alpha r(u)$ with $\alpha \sup_{v \in \mathcal{M}_{\text{base}}} \|r(v)\|_2 < \tau_0/2$:*

1. *If $r(u) = r_\perp(u) \in N_u\mathcal{M}_{\text{base}}$ for every u, then $\pi\big(z_\alpha(u)\big) = u$: normal enrichment preserves the base coordinate.*
2. *If $r(u) = r_T(u) \in T_u\mathcal{M}_{\text{base}}$, then $\pi\big(z_\alpha(u)\big) = u + \alpha r_T(u) + \mathcal{O}(\alpha^2)$ as an ambient first-order expansion in a local chart: tangential enrichment shifts it at first order.*

This asymmetry is the central result of our theoretical analysis. It shows that *not all enrichment is equal*: the direction of the added detail, not merely its magnitude, determines whether the base coordinate is preserved or shifted under tubular projection. Normal enrichment moves points along the normal fiber, leaving the base coordinate invariant; tangential enrichment mixes added detail directly into the base coordinate and thus into the transport directions that diffusion has already learned to navigate.

Together with Proposition 3.2, this yields a concrete design principle:

Normal Projection Principle. *When enriching a navigable base latent, project detail residuals onto the normal space of the base manifold before fusion:*

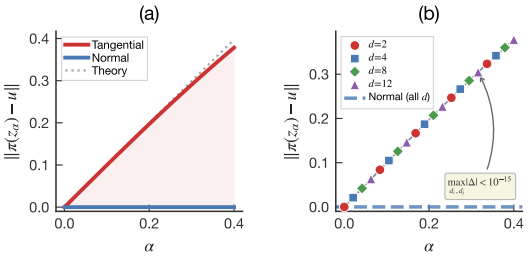


Figure 2: **Direction matters on controlled spheres.** On $S^4(1)$, normal enrichment causes essentially zero projection displacement, while tangential enrichment induces a displacement that scales linearly with α and matches $\|\alpha\, r_T\|$. The same separation holds for $d \in \{2, 4, 8, 12\}$, supporting the tubular-projection calculation in Proposition 3.3.

$$z_\alpha \;=\; u + \alpha\, \Pi_{N_u\mathcal{M}_{\text{base}}}(r). \tag{7}$$

Under this rule, $\pi(z_\alpha) = u$: added detail lives on the normal fiber rather than in the transport directions. The principle does not claim that all geometry is unchanged; it preserves the base coordinate system induced by π, which is the part most directly tied to reducing disruption of a navigable latent.

3.4 Synthetic Sanity Check

We sanity-check the Normal Projection Principle on spheres $S^d(1)$ with known geometry, using $d \in \{2, 4, 8, 12\}$. Spheres are useful because their tangent and normal spaces, nearest-point projection, and reach are all available in closed form, so the effect of residual direction can be isolated without estimator noise. We perturb a base point u with either a normal or tangential residual and measure $\|\pi(z_\alpha) - u\|$ as the fusion strength α varies; a complementary verification of Proposition 3.2 appears in Appendix E. As shown in Figure 2, normal enrichment gives essentially zero displacement, while tangential enrichment grows linearly in α, matching Proposition 3.3. This experiment is not intended to prove sample-quality gains by itself; rather, it verifies the local geometric mechanism that the model later implements with PCA-estimated token manifolds. The clean separation across dimensions provides a controlled reference point for interpreting the real-latent diagnostics in Section 5.3 and Appendix F.

4 NormFuse: From Geometric Principle to Model Design

Having established the Normal Projection Principle, we translate it into a concrete RAE-based architecture. Existing RAE-style models do not enforce this principle: RAE [Zheng et al., 2025]

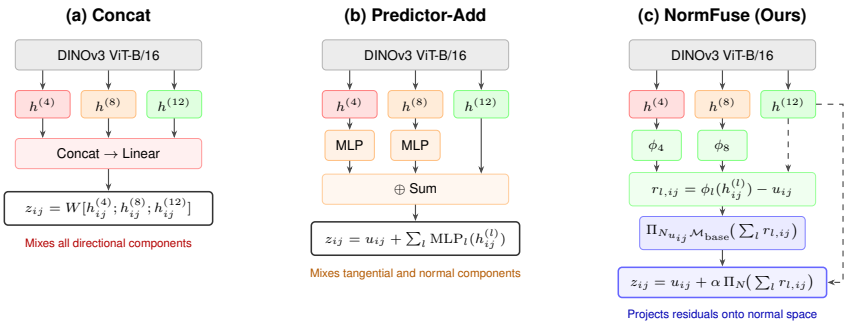


Figure 3: **Three fusion strategies for multi-layer VFM features. (a)** Concat mixes all directional components via a learned projection. **(b)** Predictor-Add sums MLP residuals that still contain both tangential and normal components. **(c)** NormFuse projects the aggregated residual onto the normal space $N_{u_{ij}}\mathcal{M}_{\mathrm{base}}$ before adding it to the base point u_{ij} (dashed = skip connection carrying u_{ij}).

uses a single encoder layer, while SVG-T2I [Shi et al., 2025] injects multi-layer detail through an unconstrained residual branch that can leak into tangential directions. NormFuse addresses both issues by fusing multi-layer DINOv3 patch-token features through token-wise normal projection; Figure 3 contrasts the three fusion strategies.

Setup. We adopt a frozen DINOv3 ViT-B/16 encoder [Siméoni et al., 2025] with patch-token dimension $D = 768$. At 512^2 resolution, each selected layer produces a spatial token grid $h^{(l)} \in \mathbb{R}^{H' \times W' \times D}$ with $H' = W' = 32$. The final-layer grid $u = h^{(12)}$ serves as the *base latent*; each token $u_{ij} \in \mathbb{R}^D$ is treated as a point on the base representation manifold. Empirically, base-layer tokens have low intrinsic dimension (estimated as $\hat{d} \approx 39$ on COCO tokens) and high local linearity (LL ≈ 0.91), making them favorable for navigability. The detail-source set $\mathcal{S} = \{4, 8, 12\}$ combines shallower detail layers with the final base layer. Layers 4 and 8 retain richer local information but exhibit less favorable geometry, while layer 12 is included for a controlled comparison with Concat and Predictor-Add and acts as a small base-coordinate anchor after alignment. Our goal is to fuse u with complementary detail from $\mathcal{S}$ into a patch-token latent grid $z \in \mathbb{R}^{H' \times W' \times D}$ that is useful for both decoding and diffusion while preserving the geometry of the base token manifold.

Fusion strategies. All fusion operations are applied token-wise with shared parameters across spatial locations. **Concat** concatenates tokens from layers $\{4, 8, 12\}$ and projects them back to D, mixing all directional components without geometric control. **Predictor-Add** maps detail tokens through per-layer MLPs and adds them to the base token u_{ij}, preserving dimensionality but allowing both tangent and normal components through; this is closest to SVG-T2I [Shi et al., 2025]. Implementation details for this baseline are given in Appendix G. **NormFuse** instead applies the Normal Projection Principle to each token residual before fusion.

NormFuse implementation. For each detail-source layer $l \in \mathcal{S}$, a learned point-wise linear map $\phi_l : \mathbb{R}^D \to \mathbb{R}^D$ aligns shallow tokens to the base-layer token space. The maps are trained independently on frozen encoder patch tokens and then fixed for decoder and diffusion training. At spatial location (i, j), the detail residual is

$$r_{l,ij} = \phi_l(h_{ij}^{(l)}) - u_{ij}. \tag{8}$$

To estimate the normal space, we build a FAISS [Johnson et al., 2019] bank of base-layer patch tokens from the frozen encoder. For each query token u_{ij}, we retrieve its $k = 64$ nearest base tokens and perform PCA on the centered neighborhood. The top components explaining 95% of local variance form $V_{d,ij} \in \mathbb{R}^{D \times d}$, which spans the estimated tangent space $\hat{T}_{u_{ij}}\mathcal{M}_{\mathrm{base}}$. The corresponding normal projector is

$$\Pi_{N_{u_{ij}}\mathcal{M}_{\mathrm{base}}} = I - V_{d,ij}V_{d,ij}^\top.$$

NormFuse then forms the fused patch-token grid as

$$z_{ij} = u_{ij} + \alpha \Pi_{N_{u_{ij}}\mathcal{M}_{\mathrm{base}}} \left(\sum_{l \in \mathcal{S}} r_{l,ij} \right), \qquad i = 1, \ldots, H',\ j = 1, \ldots, W', \tag{9}$$

where α controls enrichment strength. By construction, each added residual lies in the estimated normal space of its base token. This implements the coordinate-preserving rule in Proposition 3.3 and is intended to reduce interference with the transport directions of the base token manifold.

Architecture and cost. The fused latent grid $z \in \mathbb{R}^{32\times32\times768}$ is decoded by a lightweight convolutional decoder, and a DiT [Peebles and Xie, 2023] trained with flow matching [Lipman et al., 2023] models the fused patch-token grid. Across Concat, Predictor-Add, and NormFuse, the encoder, decoder, DiT, data, and evaluation protocol are fixed, isolating the fusion mechanism. Projectors are constructed and cached before diffusion training; at inference, the DiT samples fused latents directly and no k-NN or PCA lookup is required. Bank/index construction takes approximately 2.0 A100 GPU-hours once. Relative to RAE-Base, latent preprocessing takes $1.58\times$ the time and $1.17\times$ the peak memory; cached-latent DiT throughput and standard text-to-image latency remain unchanged to the reported precision, with throughput differences within 1%. These measurements concern preprocessing, not extra sampling steps. Full architectural details are in Appendix G.

5 Experiments

5.1 Experimental Setup

Datasets. For controlled ablations, we train on COCO 2017 [Lin et al., 2014] ($\sim$118K images at 512^2 resolution) and evaluate reconstruction on its 5K validation split. For large-scale text-to-image generation, we train on approximately 50M image–text pairs from BLIP-3o [Chen et al., 2025a] and DanQing [Shen et al., 2026], with six captions per image (short, medium, and long in Chinese and English).

Model configurations. The main experiments use the same frozen DINOv3 ViT-B/16 encoder [Siméoni et al., 2025], with patch-token dimension D=768, latent grid size 32×32 at 512^2, base layer l=12, and detail sources $\mathcal{S} = \{4, 8, 12\}$. For ablations, we pair this encoder with a DiT-S/2 backbone (12 layers, hidden size 384, 33M parameters); the scaled text-to-image model uses a 4B-parameter DiT (32 layers, hidden size 2560, 32 attention heads, 8 KV heads, M-RoPE, patch size 1) and a frozen Qwen3-4B [Yang et al., 2025] text encoder via cross-attention. Training is progressive: 256^2 for 200K steps, 512^2 for 150K steps, and 1024^2 for 50K steps on a curated 1M subset, using AdamW with cosine learning-rate decay.

Evaluation metrics. We measure reconstruction with PSNR and rFID on the COCO 2017 validation split, generation quality with gFID on 50K samples, and compositional alignment with GenEval [Ghosh et al., 2023] and DPG-Bench [Hu et al., 2024] using 50 Euler steps and classifier-free guidance scale 4.5. We also report three geometric diagnostics on 50K sampled token vectors: intrinsic dimension (ID), local linearity (LL), and normal fraction (NF).

Baselines. For controlled comparisons, we evaluate Base-only, Concat, Predictor-Add, and NormFuse under a shared encoder, decoder, and DiT backbone. For large-scale generation, matched RAE-Base and Predictor-Add baselines share the same DINOv3 encoder, decoder family, 4B DiT, Qwen3-4B text encoder, data mixture, sampler, and optimization schedule as NormFuse: RAE-Base uses only $h^{(12)}$, while Predictor-Add uses the same multi-layer sources but adds unconstrained MLP residuals without normal projection. We also compare against representative VAE-based diffusion, autoregressive, and RAE-based methods.

5.2 Main Results

Tables 1 and 2 compare NormFuse with representative systems and, most importantly, with two newly added matched large-scale baselines. RAE-Base is trained with the same 50M image–text corpus, Qwen3-4B text encoder, 4B DiT backbone, decoder family, progressive schedule, sampler, and evaluation protocol as NormFuse but uses only the final DINOv3 layer $h^{(12)}$. Predictor-Add uses the same multi-layer sources and training recipe as NormFuse, but replaces normal projection with unconstrained MLP residual addition. These two controls separate the value of multi-layer enrichment from the value of the normal-space constraint.

Matched large-scale comparison. RAE-Base shows the cost of relying only on the final VFM layer. On GenEval, NormFuse improves over RAE-Base from 0.64 to 0.77 overall, with large gains on two-object composition (from 0.66 to 0.85), colors (from 0.72 to 0.85), and position (from 0.42 to

Table 1: GenEval results. RAE-Base and Predictor-Add are matched baselines trained with the same large-scale recipe as NormFuse; † denotes an LLM prompt rewriter.

	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall ↑
VAE Diff.	SDXL [Podell et al., 2024]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [Betker et al., 2023]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [Esser et al., 2024]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	Lumina-2.0† [Qin et al., 2025]	–	0.87	0.67	–	–	0.62	0.73
	FLUX.1-dev† [Black Forest Labs, 2024]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
AR	Emu3-Gen† [Wang et al., 2024]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
	Janus-Pro-7B [Chen et al., 2025b]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
RAE	Scale-RAE-2.4B [Tong et al., 2026]	0.95	0.54	0.40	0.62	0.24	0.13	0.48
	Scale-RAE-9.8B [Tong et al., 2026]	0.97	0.68	0.45	0.66	0.24	0.25	0.54
	SVG-T2I† [Shi et al., 2025]	0.94	0.89	0.49	0.89	0.69	0.62	0.75
	RAE-Base (Ours, matched)	0.93	0.66	0.48	0.72	0.42	0.61	0.64
	Predictor-Add (Ours, matched)	0.95	0.78	0.50	0.71	0.69	0.64	0.70
	NormFuse (Ours)	0.94	0.85	0.54	0.85	0.74	0.67	0.77

Table 2: DPG-Bench results. RAE-Base and Predictor-Add are matched baselines trained with the same large-scale recipe as NormFuse.

	Method	Global	Entity	Attribute	Relation	Other	Overall ↑
VAE Diff.	SDXL [Podell et al., 2024]	83.27	82.43	80.91	86.76	80.41	74.65
	DALL-E 3 [Betker et al., 2023]	90.97	89.61	88.39	90.58	89.83	83.50
	SD3-Medium [Esser et al., 2024]	87.90	91.01	88.83	80.70	88.68	84.08
	FLUX.1-dev [Black Forest Labs, 2024]	74.35	90.00	88.96	90.87	88.33	83.84
	HiDream-I1 [Cai et al., 2025]	76.44	90.22	89.48	93.74	91.83	85.89
	Lumina-2.0 [Qin et al., 2025]	–	91.97	90.20	94.85	–	87.20
AR	Emu3-Gen [Wang et al., 2024]	85.21	86.68	86.84	90.22	83.15	80.60
	Janus-Pro-7B [Chen et al., 2025b]	86.90	88.90	89.40	89.32	89.48	84.19
RAE	Scale-RAE-2.4B [Tong et al., 2026]	77.20	81.34	75.70	76.48	69.60	71.76
	Scale-RAE-9.8B [Tong et al., 2026]	74.77	83.60	78.86	78.17	74.80	74.85
	SVG-T2I [Shi et al., 2025]	88.50	91.00	91.86	92.21	91.86	85.78
	RAE-Base (Ours, matched)	85.79	83.04	81.27	81.44	79.60	79.57
	Predictor-Add (Ours, matched)	86.33	84.17	80.55	86.31	84.73	82.87
	NormFuse (Ours)	89.45	92.26	94.03	90.58	92.44	86.89

0.74). On DPG-Bench, NormFuse raises the overall score from 79.57 to 86.89, with broad gains across Global, Entity, Attribute, Relation, and Other. This pattern suggests that the final DINOv3 layer provides a semantic and navigable base latent, but lacks the multi-layer local detail needed for reliable compositional binding.

Predictor-Add directly tests whether unconstrained multi-layer enrichment is sufficient at scale. NormFuse improves GenEval overall from 0.70 to 0.77, with especially clear gains in colors (0.71 to 0.85), two-object composition (0.78 to 0.85), and position (0.69 to 0.74). On DPG-Bench, NormFuse improves overall from 82.87 to 86.89 and raises Entity, Attribute, Relation, and Other by 8.09, 13.48, 4.27, and 7.71 points, respectively. The slight drop on single-object GenEval (0.95 to 0.94) is small relative to the gains in compositional categories. Because these baselines share the same large-scale recipe, the comparison provides matched evidence that the gains come from the latent construction and the normal-space constraint rather than from data scale or text-encoder strength.

Comparison with external baselines. Against heterogeneous published systems, NormFuse reaches 0.77 on GenEval, outperforming Scale-RAE and SVG-T2I among RAE-based methods. On DPG-Bench, it obtains 86.89, surpassing SVG-T2I by 1.11 overall and all listed autoregressive systems; the gains are in overall score rather than uniformly across every submetric. These external comparisons are used for positioning because systems differ in data, scale, prompt processing, and engineering details. The matched RAE-Base and Predictor-Add results provide the cleaner evidence for our central claim: normal-space multi-layer enrichment is stronger than both single-layer RAE latents and unconstrained residual fusion.

Together with the controlled COCO ablations in Table 3, these large-scale results support the same conclusion in two regimes: NormFuse primarily avoids the reconstruction–generation degradation of unconstrained fusion while preserving the reconstruction benefit of multi-layer detail.

Table 3: **Fusion strategy comparison.**

Method	PSNR ↑	rFID ↓	gFID ↓	ID ↓	LL ↑	NF ↑
Base only ($l=12$)	20.3	3.82	5.54	39	0.91	–
Concat ($l=4,8,12$)	26.1	**1.51**	5.91	78	0.81	0.38
Predictor-Add ($l=4,8,12$)	**27.6**	1.53	5.94	61	0.85	0.52
NormFuse ($l=4,8,12$)	27.4	1.58	**5.33**	**35**	**0.93**	**0.91**

Table 4: **Compact ablation.** Default configuration is $\mathcal{S}=\{4,8,12\}$, $\alpha=1.0$, and $k=64$.

Category	Setting	PSNR ↑	rFID ↓	gFID ↓	NF ↑
Detail sources	$\emptyset$ (base only)	20.3	3.82	5.54	–
	$\{8\}$	24.7	2.64	5.71	0.84
	$\{4,8\}$	27.2	1.81	5.61	0.87
	$\{4,8,12\}$	**27.4**	**1.58**	**5.33**	**0.91**
Enrichment strength	$\alpha=0.5$	26.0	1.96	**5.24**	**0.93**
	$\alpha=1.0$	**27.4**	1.58	5.33	0.91
	$\alpha=1.5$	**27.4**	**1.56**	5.62	0.89
PCA neighborhood	$k=32$	27.1	1.65	**5.32**	0.88
	$k=64$	**27.4**	**1.58**	5.33	**0.91**
	$k=256$	27.1	1.62	5.37	0.87

Geometry at scale and a second encoder. For the matched DINOv3/4B models, ID/LL/NF are $42/0.90/-$ for RAE-Base, $66/0.84/0.50$ for Predictor-Add, and $38/0.92/0.89$ for NormFuse. A matched SigLIPv2/4B experiment with 5M pairs improves GenEval from 0.64 to 0.70 and DPG-Bench from 79.9 to 83.2 over Predictor-Add, at similar reconstruction quality (25.9 versus 25.7 PSNR; 1.79 versus 1.82 rFID). This supports transfer at 4B capacity; it is not a second 50M-pair replication.

5.3 Ablation Studies

All ablations use the COCO 2017 setup with DiT-S/2 (33M parameters) and hold the encoder, decoder, diffusion backbone, training data, sampling protocol, and evaluation pipeline fixed, changing only the fusion rule or one NormFuse hyperparameter.

Fusion strategy comparison. Table 3 compares fusion rules under a shared encoder, decoder, and DiT backbone. Base-only gives competitive gFID but weak reconstruction, while Concat and Predictor-Add improve reconstruction at the cost of higher gFID, higher ID, and lower LL. NormFuse keeps reconstruction close to Predictor-Add (27.4 vs. 27.6 PSNR; 1.58 vs. 1.53 rFID), preserves geometry closer to Base-only, and obtains the best gFID (5.33). This is the key matched evidence for the paper's mechanism: unconstrained fusion adds useful detail but changes the latent geometry in ways associated with harder diffusion. NormFuse retains nearly the same reconstruction benefit while reducing ID from 61 to 35 relative to Predictor-Add and increasing LL from 0.85 to 0.93; its NF of 0.91 is an implementation sanity check under an independently recomputed projector. We therefore frame the strongest small-scale evidence as the 0.61 gFID gain over Predictor-Add, while the 0.21 gain over Base-only should be read as a smaller margin.

Geometry and flow-model fitting. On the same COCO/DiT-S comparison, NormFuse reduces learned-velocity roughness by 22% and held-out flow-matching loss by 8.5% relative to Predictor-Add, whose normalized values are 1.00 and 1.000 (NormFuse: 0.78 and 0.915). Evaluation shares base-latent PCA projectors, validation images, flow-time samples, noise draws, seeds, and normalization. Across 12 unique controlled configurations, Spearman correlations are 0.84 for roughness–loss, 0.69 for roughness–gFID, and 0.76 for loss–gFID. These are learned-field associations, not direct estimates of $\mathcal{C}_T$ or quantitative FID guarantees.

Component ablation and hyperparameter sensitivity. Table 4 shows that adding layers 8 and 4 improves reconstruction while keeping gFID close to Base-only (5.61 vs. 5.54); adding layer 12 further improves gFID to 5.33, with residual calibration as a possible, not separately isolated, explanation (Appendix G). For enrichment strength, $\alpha=0.5$ gives the lowest gFID (5.24), while

$\alpha = 1.0$ improves PSNR by 1.4 dB and rFID by 0.38 with a small gFID cost; we use $\alpha = 1.0$ as the default compromise. The degradation at $\alpha = 1.5$ and the lower NF are consistent with, but do not independently establish, the tubular-projection intuition: overly large residuals are more likely to leave the local regime where the estimated normal fiber is reliable. For PCA neighborhood size, all three settings remain close (gFID 5.32–5.37), with $k = 64$ giving the highest NF and best reconstruction; the default $\alpha = 1.0, k = 64$ is therefore chosen as a balanced configuration rather than the single best point for any one metric.

6 Conclusion

We presented a geometric view of latent diffusion through the tension between invertibility and navigability. Our analysis identifies intrinsic dimension, curvature, reach, and enrichment direction as useful factors for latent design: normal residuals preserve the base coordinate, while tangential residuals shift it at first order. NormFuse instantiates this principle by fusing multi-layer DINOv3 features through token-wise normal projection. Controlled ablations show that it preserves strong reconstruction without the generation degradation of unconstrained fusion, and matched large-scale comparisons show clear gains over both final-layer RAE-Base and unconstrained Predictor-Add baselines. These results support normal-space enrichment as a practical latent-design rule for RAE-based text-to-image generation.

References

Sumukh K. Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems*, 2024.

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. OpenAI technical report, 2023. https://cdn.openai.com/papers/dall-e-3.pdf.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. HiDream-I1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, Emile Mathieu, Michael J. Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR, 2024.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. Introduces DPG-Bench.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 28795–28831. PMLR, 2025.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 31648–31666. PMLR, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

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

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Yuan, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-Image 2.0: A unified and efficient image generative framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Hengyu Shen, Tiancheng Gu, Bin Qin, Lan Wu, Yuling Wu, Shuo Tan, Zelong Sun, Jun Wang, Nan Wu, Xiang An, Weidong Cai, Ziyong Feng, and Kaicheng Yang. DanQing: An up-to-date large-scale Chinese vision-language pre-training dataset. *arXiv preprint arXiv:2601.10305*, 2026.

Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. SVG-T2I: Scaling up text-to-image latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2512.11749*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 55876–55905. PMLR, 2025.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Haotian Zhang, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction FID predictive of diffusion generation FID. *arXiv preprint arXiv:2603.05630*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger

Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimension reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Supplementary Material for
The Geometry of Generative Latents: Navigability and Invertibility

A Overview

This supplementary material provides formal proofs, extended empirical analyses, and full implementation details that support the main manuscript. The document is organized as follows:

- **Section B**: Proof of Proposition 3.2, which bounds the tubular tangent-projected contribution of a smooth data manifold by its intrinsic dimension, curvature, and reach.
- **Section C**: Proof of Proposition 3.3, establishing the directional asymmetry between normal and tangential enrichment under tubular projection.
- **Section D**: Cross-encoder generality of the Normal Projection Principle on three ViT-B vision foundation models (DINOv2, SigLIPv2, and MAE).
- **Section E**: Synthetic verification of the tangent-projected transport complexity bound on a manifold zoo of spheres, tori, flat compact patches, and product manifolds.
- **Section F**: Geometric analysis of real learned latent spaces, connecting intrinsic dimension and normal fraction to generation quality.
- **Section G**: Implementation and training details for both the controlled ablations and the scaled text-to-image experiments.
- **Section H**: Extended discussion of related work, covering manifold learning, score-based diffusion on manifolds, and representation autoencoders.
- **Section I**: Additional qualitative showcases of NormFuse across diverse prompts and visual styles.
- **Section J**: Broader impacts and asset-license information.
- **Section K**: Current limitations of the framework across theoretical scope, implementation choices, and scaling.
- **Section L**: Three concrete directions for future work.

B Proof of Proposition 3.2

We prove the qualitative bound for the tangent-projected transport complexity in Definition 3.1. The projection onto $T_{\pi(z)}\mathcal{M}$ is important: the raw ambient Hessian contains a universal Gaussian normal term that scales with codimension even for a flat manifold. Since this term is not a property of the transport geometry along the data support, the main paper analyzes the tangent-projected Hessian.

Setup and tubular coordinates. Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact C^3 d-dimensional embedded submanifold with reach $\tau > 0$ and sectional curvature bounded by K. For z in the tubular neighborhood $\mathcal{T}_\tau(\mathcal{M})$, write $z = u + \nu$, where $u = \pi(z) \in \mathcal{M}$ and $\nu \in N_u\mathcal{M}$. Let $\Pi_T(z) = \Pi_{T_u\mathcal{M}}$. The reach condition makes this decomposition unique and smooth, and bounds the operator norm of the second fundamental form by a constant of order $1/\tau$.

Local expression of the smoothed density. Under the Gaussian forward process, $p_t = p * \mathcal{N}(0, tI)$. For $t \in [t_{\min}, T]$ and $z = u + \nu$ in the tubular neighborhood, Laplace expansion in the tangent coordinates gives

$$\log p_t(z) = -\frac{\|\nu\|^2}{2t} + \log q_t(u, \nu) + \mathrm{const}(t, D - d),$$

where q_t contains the on-manifold density, the Jacobian of the tubular coordinate map, and curvature-dependent corrections. The first term is the universal normal Gaussian contribution. For a flat embedded manifold, its leading ambient Hessian is purely normal and vanishes under $\Pi_T(\cdot)\,\Pi_T$. For a curved manifold and off-manifold noisy samples, however, the squared-distance Hessian also contains curvature-dependent tangent terms through the variation of the normal frame. These terms

are not universal codimension terms; they are controlled by the second fundamental form and are absorbed into the curvature- and reach-dependent part of the bound.

Tangent Hessian control. In a local orthonormal tangent frame, the projected Hessian decomposes into three contributions: (i) the intrinsic Hessian of the smoothed on-manifold density, (ii) curvature corrections induced by the second fundamental form and frame variation, and (iii) finite-noise smoothing terms controlled by $t_{\min}$ and T. The density assumptions give uniform bounds on the first term. The curvature and reach assumptions give

$$\|\nabla_T^2 \log q_t\|_F^2 \le C_1 d^2 \left(1 + \frac{K}{\tau^2} + \frac{1}{\tau^4}\right)$$

uniformly for $t \in [t_{\min}, T]$, where C_1 depends on the density bounds, their first two derivatives, the finite noise interval, and the normalization scale of the latent coordinates. The factor d^2 comes from the number of tangent Hessian entries under uniform per-entry control, while the curvature and reach terms control how rapidly the tangent frame, the normal frame, and the projection map can vary. For manifolds whose mean curvature grows with d, additional dimension-dependent constants can enter this qualitative envelope. The expression is written in normalized latent units; equivalently, the dimensional constants associated with the finite noise interval are absorbed into C_1.

Tail outside the tubular neighborhood. For small and moderate noise levels, the mass of p_t outside $\mathcal{T}_\tau(\mathcal{M})$ is controlled by the Gaussian tail relative to the reach and the effective codimension. In high ambient dimension, isotropic noise can place non-negligible mass outside the tubular neighborhood unless the effective training-noise range is sufficiently local. We therefore keep this contribution explicit as R_{tail} in Definition 3.1 and do not interpret the displayed tubular bound as a codimension-uniform guarantee for arbitrary noise scales. For larger $t \le T$, smoothness of the Gaussian convolution gives a bounded residual contribution that is absorbed into this tail term.

Integration over the finite training-noise range. Integrating over $t \in [t_{\min}, T]$ yields

$$\mathcal{C}_T(\mathcal{Z}) - R_{\text{tail}} \le C d^2 \left(1 + \frac{K}{\tau^2} + \frac{1}{\tau^4}\right),$$

where C depends on the density regularity, $\mathrm{vol}(\mathcal{M})$, $t_{\min}$, T, and the normalization scale of the latent coordinates. The bound should be interpreted as a structural bound on the tubular tangent-projected contribution; R_{tail} records the off-tube residual and may depend on the effective codimension and noise range. This proves the qualitative form of Proposition 3.2. □

C Proof of Proposition 3.3

We prove the proposition as a statement about *tubular projection* rather than as a quantitative bound on tangent-projected transport complexity. The goal is to show that normal and tangential enrichment interact with the base manifold in structurally different ways.

Setup. Let $\psi : U \subset \mathbb{R}^d \to \mathcal{M}_{\text{base}}$ be a local chart around $u = \psi(y)$, and decompose the residual field as

$$r(\psi(y)) \;=\; D\psi_y\, w(y) + \eta(y), \qquad D\psi_y\, w(y) \in T_{\psi(y)}\mathcal{M}_{\text{base}}, \qquad \eta(y) \in N_{\psi(y)}\mathcal{M}_{\text{base}}, \quad (10)$$

so that $r_T(\psi(y)) = D\psi_y\, w(y)$ and $r_\perp(\psi(y)) = \eta(y)$. We assume $\alpha \sup_{v \in \mathcal{M}_{\text{base}}} \|r(v)\|_2 < \tau_0/2$, so every enriched point lies in the tubular neighborhood on which the nearest-point projection π is well defined and smooth.

Step 1: Pure normal enrichment preserves the base coordinate. For $r = r_\perp$, the augmented point is $z_\alpha(u) = u + \alpha r_\perp(u)$ with $r_\perp(u) \in N_u\mathcal{M}_{\text{base}}$. By definition of the tubular neighborhood, any point of the form $u + v$ with $v \in N_u\mathcal{M}_{\text{base}}$ and $\|v\| < \tau_0$ lies on the normal fiber over u, so

$$\pi\big(z_\alpha(u)\big) \;=\; u. \quad (11)$$

Pure normal enrichment therefore changes the ambient embedding while preserving the identity of the underlying base latent under tubular projection.

Step 2: Pure tangential enrichment shifts the base coordinate at first order. For $r = r_T$, a first-order Taylor expansion of the chart gives

$$\psi(y + \alpha w(y)) \;=\; \psi(y) + \alpha D\psi_y w(y) + \mathcal{O}(\alpha^2), \tag{12}$$

so $z_\alpha(\psi(y)) = \psi(y) + \alpha D\psi_y w(y)$ differs from the nearby manifold point $\psi(y + \alpha w(y))$ only by a second-order term. Since π is smooth on the tubular neighborhood and acts as the identity on $\mathcal{M}_{\text{base}}$,

$$\pi\big(z_\alpha(u)\big) \;=\; u + \alpha r_T(u) + \mathcal{O}(\alpha^2), \tag{13}$$

where the equality is understood as an ambient first-order expansion in the local chart. A tangential perturbation thus shifts the projected base coordinate already at first order.

Step 3: Mixed residuals. For a general residual $r = r_T + r_\perp$, the normal component moves the point along the normal fiber while the tangential component shifts the projected base coordinate at first order, so the first-order change in $\pi(z_\alpha(u))$ is governed entirely by r_T. This is the sense in which tangential leakage entangles added detail with the transport directions of the base latent, whereas normal-aligned enrichment preserves those directions.

Remark. The proposition does not claim that normal enrichment leaves all geometric quantities unchanged, since a normal graph over $\mathcal{M}_{\text{base}}$ can still alter the induced metric and curvature. The structural point is narrower and more relevant for model design: under normal enrichment the base coordinate is preserved under tubular projection, whereas under tangential enrichment it shifts at first order. This is the property invoked in Section 3.3 to motivate projecting residuals onto the normal space before fusion. □

D Generality of the Reconstruction–Generation Trade-off Across Vision Foundation Models

The controlled ablation in Section 5.3 is conducted on a frozen DINOv3 ViT-B/16 encoder. A natural question is whether the observed patterns are specific to DINOv3 or reflect a more general property of fusing multi-layer features from strong visual representations. This appendix tests the latter hypothesis by repeating the fusion strategy comparison on three additional frozen vision foundation models of comparable capacity: DINOv2 ViT-B/14 [Oquab et al., 2024], SigLIPv2 ViT-B/16 [Tschannen et al., 2025], and MAE ViT-B/16 [He et al., 2022]. The three encoders span the main families of ViT-B visual pretraining, namely self-distillation (DINOv2), image–text contrastive alignment (SigLIPv2), and masked image modeling (MAE).

Experimental setup. For each encoder, we rebuild the entire pipeline from scratch while keeping the encoder frozen. We train the per-layer alignment maps ϕ_l on ImageNet-1K [Deng et al., 2009] patch tokens following the protocol in Appendix G, build a FAISS feature bank of base-layer patch tokens for token-wise tangent-space estimation, and train the decoder and a DiT-S/2 diffusion backbone for 400K steps at batch size 256 on COCO 2017 at 512^2. The base layer is the final transformer block ($l = 12$) in all three encoders, and the detail-source layers are $\mathcal{S} = \{4, 8, 12\}$ across all fusion strategies, matching the main paper setup. When an encoder has a different native patch grid, such as DINOv2 ViT-B/14, we bilinearly resample its patch-token grid to the common 32×32 interface before decoder and DiT training. All metrics (PSNR, rFID, gFID, ID, LL, NF) are then computed exactly as in Section 5.1. Consequently, the only variable across rows within each encoder group is the fusion strategy, and the only variable across encoder groups is the frozen backbone.

Results. Table 5 shows that the patterns documented on DINOv3 reappear under every encoder. Concat and Predictor-Add improve reconstruction over the Base-only condition ($+5.7$ to $+7.1$ dB PSNR and rFID reductions of roughly $2\times$ to $3\times$) while modestly worsening gFID ($+0.2$ to $+1.0$), and the geometric diagnostics inflate in the same direction (ID by $1.4\times$ to $2.3\times$, LL drops by 0.09 to 0.15, NF at most 0.50 for Predictor-Add and below 0.40 for Concat). NormFuse, in contrast, matches Predictor-Add on reconstruction within 0.1 to 0.2 dB PSNR while keeping ID and LL close to the base values and NF high. It is the lowest-gFID fusion strategy on all three encoders; compared with the Base-only reference, it improves on DINOv2 (5.82 vs. 5.93) and SigLIPv2 (6.22 vs. 6.28), and remains within 0.08 on MAE (6.54 for NormFuse vs. 6.46 for Base-only). The gFID gap between NormFuse and Predictor-Add is 0.35 on DINOv2, 1.01 on SigLIPv2, and 0.33 on MAE.

Table 5: **Cross-encoder fusion strategy comparison.** All runs use the same COCO/DiT-S protocol; best values within each encoder block are bold.

		Reconstruction		Generation	Geometry		
Encoder	Fusion strategy	PSNR ↑	rFID ↓	gFID ↓	ID ↓	LL ↑	NF ↑
DINOv2	Base-only ($l=12$)	19.1	3.65	5.93	**36**	**0.93**	–
	Concat ($\mathcal{S}$)	24.8	**1.68**	6.20	82	0.80	0.36
	Predictor-Add ($\mathcal{S}$)	**26.2**	**1.68**	6.17	65	0.84	0.50
	NormFuse ($\mathcal{S}$)	26.0	1.70	**5.82**	39	0.92	**0.90**
SigLIPv2	Base-only ($l=12$)	18.6	4.94	6.28	**38**	**0.92**	–
	Concat ($\mathcal{S}$)	24.3	**1.80**	7.11	86	0.78	0.34
	Predictor-Add ($\mathcal{S}$)	**25.7**	1.82	7.23	67	0.83	0.49
	NormFuse ($\mathcal{S}$)	25.5	1.84	**6.22**	41	0.91	**0.89**
MAE	Base-only ($l=12$)	18.9	5.38	**6.46**	**58**	**0.87**	–
	Concat ($\mathcal{S}$)	24.6	1.73	6.90	104	0.72	0.30
	Predictor-Add ($\mathcal{S}$)	**25.9**	**1.70**	6.87	82	0.78	0.45
	NormFuse ($\mathcal{S}$)	25.8	1.78	6.54	61	0.86	**0.88**

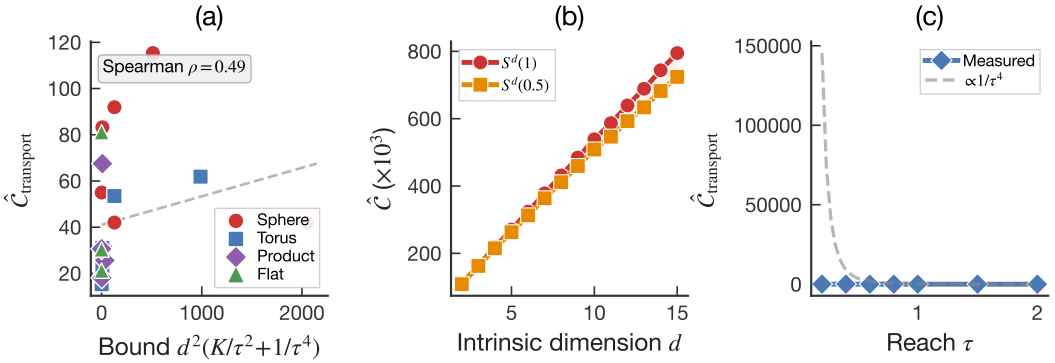


Figure 4: **Manifold Zoo verification of the tangent-projected transport complexity bound. (a)** Empirical tangent-projected transport complexity versus the curvature–reach proxy $d^2(K/\tau^2+1/\tau^4)$ across four manifold families (sphere, torus, flat compact patch, product). This visualization omits the additive constant in Eq. 4 to emphasize geometry-dependent variation; the monotone trend (Spearman $\rho = 0.49$) captures the coarse ordering despite estimator noise. **(b)** $\mathcal{C}_T$ scales quadratically with intrinsic dimension d on spheres of radius $r=1$ and 0.5, matching the predicted d^2 dependence. **(c)** Decreasing reach τ (smaller sphere radius) sharply increases tangent-projected transport complexity, consistent with the $1/\tau^4$ term.

Takeaway. Three encoder families, spanning self-distillation, image–text contrastive, and masked image modeling, yield the same qualitative pattern: unconstrained fusion trades generation for reconstruction, whereas normal-projected fusion mitigates this trade-off with at most a negligible reconstruction penalty (≤ 0.3 dB PSNR and ≤ 0.08 rFID across all three encoders). Absolute generation quality still varies with the navigability of the base latent (MAE has the highest Base ID at 58 and the worst gFID, DINOv2 the lowest at 36 and the best gFID), consistent with the diagnostic role of Proposition 3.2. The geometric view proposed in this paper is therefore not a DINOv3-specific observation, although the exact ID shift induced by NormFuse is encoder-dependent.

E Synthetic Verification of the Tangent-Projected Transport Complexity Bound

Section 3.4 verifies Proposition 3.3 (directional asymmetry) on synthetic spheres. This appendix provides a parallel verification of Proposition 3.2 (the tangent-projected transport complexity bound) on a broader family of manifolds with known geometry.

Manifold families. We evaluate the bound on four families of smooth compact submanifolds for which the intrinsic dimension d, the sectional curvature bound K, and the reach τ are known in closed form:

- **Spheres** $S^d(r) \subset \mathbb{R}^{d+1}$: intrinsic dimension d, constant sectional curvature $K = 1/r^2$, and reach $\tau = r$. Varying d and r independently isolates dimension and reach effects.
- **Tori of revolution** $T^2(R, r) \subset \mathbb{R}^3$: intrinsic dimension $d = 2$, with curvature and reach controlled by the two radii. This family provides a non-simply-connected test case with mixed curvature.
- **Flat compact patches** $[-1, 1]^d \subset \mathbb{R}^D$: zero curvature in the interior and effectively large reach away from the boundary. These serve as a Euclidean baseline for minimal tangent-projected transport complexity; they are included as a controlled diagnostic baseline rather than as a compact boundaryless manifold satisfying all assumptions of Proposition 3.2.
- **Product manifolds** $S^{d_1}(r_1) \times S^{d_2}(r_2)$: intrinsic dimension $d_1 + d_2$, with geometry inherited from both factors. This family tests how the bound behaves under product structure.

Protocol. For each manifold instance we sample points from its natural uniform distribution; for the flat compact patches this means the uniform distribution on the bounded patch. We then apply the standard Gaussian forward process, estimate the score with a small neural network, and compute the empirical tangent-projected transport complexity by numerically integrating $\mathbb{E}[\|\Pi_T \nabla_z^2 \log p_t(z) \Pi_T\|_F^2]$ over the diffusion time interval. This procedure isolates the effect of manifold geometry on diffusion difficulty, while holding the score estimator and the evaluation protocol fixed across all instances.

Cross-family ordering (Figure 4a). The monotone trend (Spearman $\rho = 0.49$) suggests that the geometric proxy captures a coarse *ordering* of diffusion difficulty across geometrically diverse manifolds. Flat compact patches are easiest, followed by large spheres, then tori and product manifolds with smaller radii. The moderate correlation, rather than a tight linear fit, reflects two sources of slack. First, the bound is qualitative, since the constant C is not tight and subsumes density-dependent factors. Second, the empirical tangent-projected transport complexity is estimated from a learned score network, which introduces noise that is largest for high-curvature manifolds.

Quadratic dimension scaling (Figure 4b). When the sphere radius is fixed and only d varies, the tangent-projected transport complexity scales as d^2, matching the quadratic prediction of Eq. 4. The same scaling holds at $r = 0.5$ (higher curvature), shifted upward by the larger K/τ^2 and $1/\tau^4$ contributions. Intrinsic dimension is therefore a prominent factor in these controlled sphere experiments, although cross-family comparisons also depend on curvature, reach, density, and estimator noise.

Inverse reach scaling (Figure 4c). Shrinking the sphere radius, and hence $\tau = r$, at fixed d sharply increases tangent-projected transport complexity, consistent with the $1/\tau^4$ term in the bound. Manifolds with small reach, meaning sharp curvature or near-proximity of distant parts in ambient space, impose substantially harder transport problems for diffusion.

Together, these three panels support the structural dependence of tangent-projected transport complexity on manifold geometry, while leaving room for estimator noise and looseness in the qualitative bound. Appendix F next asks whether the same geometric quantities predict generation quality on *real* learned latent spaces.

F Geometric Analysis of Learned Latent Spaces

Our framework suggests two diagnostic patterns for real learned latent spaces: (i) higher intrinsic dimension should be associated with harder diffusion under matched conditions (Proposition 3.2), and (ii) enrichment aligned with the normal space of the base manifold should preserve generation quality better than tangential leakage (Proposition 3.3). We examine these patterns using the four conditions from Table 3, which share the same encoder, decoder, and diffusion architecture and differ only in how multi-layer features are combined. This controlled setup is designed to isolate the fusion mechanism, while gFID is used as an empirical outcome rather than a direct measurement of $\mathcal{C}_T$.

F.1 Large-scale Matched Baselines

The matched RAE-Base and Predictor-Add baselines are designed to remove the main confounders in external comparisons. Both use the same data mixture, Qwen3-4B text encoder, 4B DiT backbone, decoder family, progressive training schedule, sampler, and evaluation protocol as NormFuse. RAE-Base removes the multi-layer residual branch and uses only the final DINOv3 feature grid $h^{(12)}$, while Predictor-Add uses the same source layers as NormFuse but adds MLP residuals without normal projection.

The main text reports the full metric breakdown. We include this subsection to emphasize the experimental control: unlike comparisons to external systems, these rows change only the latent construction. RAE-Base serves as the large-scale counterpart of the Base-only condition, while Predictor-Add directly tests whether unconstrained multi-layer enrichment is sufficient at scale. NormFuse improves over both, supporting the conclusion that a navigable final-layer latent benefits most from geometry-preserving enrichment.

F.2 Intrinsic Dimension Correlates with Generation Difficulty

If the tangent-projected transport complexity bound is a useful diagnostic for learned latent spaces, then fusion strategies that inflate intrinsic dimension should tend to make generation harder under matched training conditions. Across the four conditions, gFID broadly tracks intrinsic dimension, and the two lowest-ID configurations also attain the lowest gFID:

- **NormFuse** (ID $= 35$, gFID $= 5.33$): normal projection adds detail while keeping ID close to the DINOv3 base value, in this case slightly below it, and yields the best generation quality among the four.
- **Base-only** (ID $= 39$, gFID $= 5.54$): the final encoder layer produces a low-dimensional, navigable latent, but sacrifices reconstruction quality.
- **Predictor-Add** (ID $= 61$, gFID $= 5.94$): unconstrained residual addition inflates the intrinsic dimension by roughly $1.6\times$ over the base and yields the weakest generation quality of the four.
- **Concat** (ID $= 78$, gFID $= 5.91$): mixing all layer features inflates ID the most and produces a gFID essentially indistinguishable from Predictor-Add, and noticeably above both NormFuse and the Base-only reference.

We read this as an ordering result rather than a numerical prediction, since the sample size is small and all four conditions share the same encoder family: the two low-ID configurations (Base-only and NormFuse) form one cluster with lower gFID, and the two high-ID configurations (Predictor-Add and Concat) form another cluster with clearly higher gFID. This ordering on real latent spaces is qualitatively consistent with the quadratic scaling observed on synthetic spheres (Figure 4b): fusion strategies that inflate intrinsic dimension are associated with worse generation quality, regardless of reconstruction performance.

NormFuse keeps ID close to (and in fact slightly below) the base value (35 vs. 39), consistent with the fact that the normal projection removes the estimated tangential component of the residual. We interpret this as a diagnostic pattern rather than a proof of causality: removing tangent-aligned residual magnitude keeps the fused tokens closer to the low-dimensional base structure while still adding reconstruction-relevant detail.

F.3 Normal Fraction Helps Explain the Reconstruction–Generation Trade-off

The intrinsic dimension analysis above helps explain why some fusion strategies produce harder latent spaces. It does not, however, explain why two methods with similar reconstruction quality can differ substantially in generation quality. We therefore turn to the *normal fraction* $\mathrm{NF} = \|r_\perp\|/\|r\|$, which measures the fraction of the added detail aligned with the normal space of the base manifold.

Predictor-Add and NormFuse form a particularly informative comparison: they achieve similar reconstruction quality (PSNR 27.6 vs. 27.4, rFID 1.53 vs. 1.58) yet differ by 0.61 in gFID (5.94 vs. 5.33). The reconstruction similarity indicates that both methods extract a comparable amount of useful detail from the shallower layers, so the generation gap is plausibly associated with *how* that detail is incorporated into the latent space. We focus on this pair rather than a full NF–gFID

regression because Concat additionally carries a very different ID (78 vs. 35–61), so its gFID is confounded with the ID effect discussed in Section F.2; comparing Predictor-Add and NormFuse better isolates the directional effect predicted by Proposition 3.3.

The normal fraction is first an implementation sanity check. Under the independently recomputed evaluation projector, NormFuse keeps 91% of the residual magnitude in the estimated normal space ($\mathrm{NF} = 0.91$), whereas Predictor-Add has substantial tangential leakage and a normal fraction of 0.52. The tangential component shifts the base coordinate under tubular projection at first order, so Predictor-Add mixes a large part of the added detail into the transport directions of the base latent, which is expected to increase tangent-projected transport complexity without providing a proportional reconstruction benefit.

These patterns support the central insight of our framework. The reconstruction–generation dilemma is not solely a capacity trade-off. It can be mitigated by controlling the *direction* of enrichment rather than merely its magnitude. Normal-space projection preserves the base coordinate system that diffusion has learned to navigate, whereas tangential leakage can disrupt it. The 0.61 gFID gap between Predictor-Add and NormFuse provides controlled evidence that this directional distinction matters at matched reconstruction quality, while the 0.21 gain over Base-only should be read as a smaller margin.

These empirical patterns on real latent spaces are consistent with the synthetic verifications: the directional asymmetry established on spheres (Section 3.4) and the dimension-dependent scaling observed on the Manifold Zoo (Appendix E) both appear in learned representations from a frozen DINOv3 encoder.

F.4 How to Interpret the Geometry Diagnostics

The diagnostics in Table 3 have different evidential roles. ID and LL are descriptive properties of the learned latent distribution and are not explicitly optimized by any fusion baseline. They are therefore useful for checking whether a fusion strategy inflates or preserves the local structure of the base latent. NF is different: because NormFuse explicitly projects residuals before fusion and NF is measured with an independently recomputed evaluation projector, a high NF mainly verifies that the intended directional constraint remains visible under evaluation rather than serving as a fully independent predictor. We therefore do not use NF alone as evidence that generation quality should improve. The more informative pattern is the combination of similar reconstruction quality, lower ID, higher LL, and better gFID for NormFuse relative to Predictor-Add.

These diagnostics should also be read as controlled evidence rather than a full causal model of sample quality. They are computed from local neighborhoods in frozen DINOv3 token space, whereas the diffusion objective optimizes a flow-matching loss over full latent grids. Thus, the diagnostics are best viewed as geometry-side probes of a failure mode: unconstrained detail injection can move residual magnitude into transport directions and inflate the effective latent geometry. The matched ablation in Table 3 tests whether avoiding that failure mode helps under fixed encoder, decoder, DiT backbone, data, and evaluation protocol.

G Implementation and Training Details

This appendix collects the implementation details needed to reproduce our experiments. Unless otherwise stated, the setup below describes the main DINOv3 experiments; Appendix D describes the cross-encoder variants.

G.1 Encoder and Feature Extraction

We use the publicly available DINOv3 ViT-B/16 checkpoint (86M parameters) as the frozen backbone encoder. Input images are resized to 512×512 and normalized with ImageNet mean and standard deviation before encoding. The encoder is frozen throughout all experiments, so no gradient flows through it during training of the decoder, the alignment maps, or the diffusion model. Each transformer layer produces a 32×32 grid of patch tokens at 512^2 input. For every selected layer l, we keep the full patch-token grid $h^{(l)} \in \mathbb{R}^{32\times32\times768}$; no global pooling is used. The feature bank, tangent-space estimation, decoder, and diffusion model all operate on patch tokens. The geometric analysis in

Section 3 is applied to each token vector in $\mathbb{R}^{768}$, while the full image latent is the spatial collection of these token-level coordinates.

G.2 Layer Selection and Alignment

The base latent is the final transformer-layer token grid ($l = 12$) of the 12-layer ViT-B/16. We write $u_{ij} = h_{ij}^{(12)}$ for the base token at spatial location (i, j). Detail token grids are drawn from $\mathcal{S} = \{4, 8, 12\}$ for all three fusion strategies (Concat, Predictor-Add, and NormFuse), so any performance difference is attributable solely to the fusion mechanism. Within $\mathcal{S}$, layers 4 and 8 exhibit large normal fractions ($\|r_\perp\|/\|r\| > 0.6$) and substantially improve reconstruction PSNR over the base alone. Layer 12 is retained so that all fusion strategies share the same source set and so that the residual branch contains an anchor in the base-layer coordinate system. Layers closer to the base (e.g., $l = 10$ and $l = 11$) produce predominantly tangential residuals and are excluded. The framework is agnostic to this specific choice, and different VFM architectures may have different optimal combinations.

For each $l \in \mathcal{S}$, a learned linear map $\phi_l : \mathbb{R}^{768} \to \mathbb{R}^{768}$ aligns each source token to the base token space. The same ϕ_l is shared over all spatial positions. We train each ϕ_l independently on frozen encoder patch tokens by minimizing

$$\min_{\phi_l} \mathop{\mathbb{E}}_{(x,i,j)} \left[\left\| \phi_l(h_{ij}^{(l)}(x)) - h_{ij}^{(12)}(x) \right\|_2^2 \right], \tag{14}$$

using AdamW (learning rate 1×10^{-3}, weight decay 0.01, batch size 2048 token samples) for 10K steps on the ImageNet training split. The maps are then frozen for all subsequent decoder and diffusion training.

Role of layer 12. For $l = 12$, ϕ_{12} converges close to the identity, so $r_{12,ij}$ is small in magnitude. Its role is mainly calibration: it anchors the summed residual branch to the base-layer feature scale and coordinate system, which can reduce the distributional bias of using only shallower residuals. Although $r_{12,ij}$ is small, including layer 12 changes the calibration of the summed residual before projection and stabilizes the residual distribution around the base feature scale. We therefore interpret the gFID improvement from adding layer 12 in Table 4 as a residual-branch anchoring effect rather than as independent semantic detail from layer 12. The detail residual is defined token-wise as $r_{l,ij} = \phi_l(h_{ij}^{(l)}) - h_{ij}^{(12)}$.

For the Predictor-Add baseline, each source layer uses a two-layer MLP with hidden width $4D$, GELU activation, and output dimension D; the output residuals are summed and added to u_{ij}. This gives Predictor-Add at least as much residual-branch capacity as the linear alignment used by NormFuse, so the controlled comparison does not favor NormFuse through a larger predictor.

G.3 Tangent Space Estimation

We estimate the tangent space of the base token manifold $\mathcal{M}_{\text{base}}$ through local PCA on a precomputed patch-token feature bank. We encode the ImageNet training set with the frozen encoder and store base-layer patch tokens in a FAISS index. In practice we use a fixed-size bank of 10M token vectors sampled uniformly over images and spatial locations, rather than image-level pooled features, and keep the same bank construction protocol across all comparisons. ImageNet is chosen as a neutral, broad-coverage feature bank that is independent of any downstream training distribution (COCO, BLIP-3o, or DanQing). Because tangent spaces are estimated from nearest neighbors in the space of *DINOv3 patch-token features* rather than in pixel space, semantically and locally similar patches from different images map to neighboring feature vectors, and the estimated projectors are stable across downstream distributions. We verified this empirically by rebuilding the bank from a COCO token subset and observing that the average subspace angle between the two PCA tangent estimates is below 3°, with downstream gFID differences under 0.05. For a given base token u_{ij}, we retrieve its $k = 64$ nearest neighboring base tokens $\mathcal{N}_k(u_{ij})$ and perform PCA on the centered neighbor set $\{u' - \bar{u} : u' \in \mathcal{N}_k(u_{ij})\}$. The tangent dimension d is chosen per token as the smallest number of components capturing at least 95% of the local variance. The normal projector is $\Pi_{N_{u_{ij}}\mathcal{M}_{\text{base}}} = I - V_{d,ij}V_{d,ij}^\top$, where $V_{d,ij} \in \mathbb{R}^{768\times d}$ contains the top d eigenvectors.

Table 6: **PCA variance-threshold sensitivity.** Threshold sweep for the local PCA tangent estimator under the COCO/DiT-S setup.

Variance threshold	PSNR ↑	rFID ↓	gFID ↓	ID ↓	LL ↑	NF ↑
90%	27.3	1.60	5.43	33	0.94	0.90
95%	27.4	1.58	5.33	35	0.93	0.91
99%	27.5	1.55	5.47	41	0.90	0.84

This estimator is heuristic, as local PCA in high-dimensional token space is only an approximation to the tangent bundle of the base representation manifold. For this reason, all controlled comparisons use the same feature bank, neighbor definition, variance threshold, and projection code, so that the main ablation changes the fusion rule rather than the estimator. Table 4 reports the main sensitivity check over k: $k = 32$, 64, and 256 give similar gFID values (5.32–5.37) and NF values (0.87–0.91), indicating that the effect is not tied to a single neighborhood size. We additionally summarize a variance-threshold sweep in Table 6. The trend is stable across 90%, 95%, and 99%: lower thresholds are slightly more conservative, while higher thresholds improve reconstruction marginally but allow more tangential leakage and worse gFID. We use 95% as the balanced default, matching the main NormFuse setting. We do not claim that the PCA projector recovers the exact tangent space; the method only requires a stable estimate of the dominant local variation directions so that residual magnitude aligned with those directions can be suppressed.

Additional local-PCA reliability and adaptive enrichment. Confidence is computed from base-token neighborhood radius and eigengap before applying either fusion rule. For four equally sized validation-image quartiles, from highest to lowest confidence, bootstrap subspace angles are 1.5°, 2.5°, 4.2°, and 7.8°; mean selected PCA ranks are 29, 35, 44, and 59. The corresponding NF values are 0.95, 0.93, 0.90, and 0.86, PSNR values are 27.8, 27.6, 27.3, and 27.0, and NormFuse/Predictor-Add flow-matching-loss ratios are 0.86, 0.89, 0.94, and 0.99. The benefit approaches parity in less reliable neighborhoods. Local PCA rank is distinct from dataset-level ID, and per-quartile FID is omitted because the subsets are small.

An exploratory confidence-adaptive variant has mean $\alpha = 0.82$. Relative to a fixed-$\alpha = 0.82$ control, it changes rFID from 1.64 to 1.63, gFID from 5.31 to 5.27, and NF from 0.91 to 0.92, with the same 27.2 PSNR. This small gFID margin is not treated as a statistically established gain. Density and eigengap are reliability indicators, not direct curvature estimates; fixed $\alpha = 1$ remains the primary setting.

G.4 Geometric Diagnostics

The three diagnostics ID, LL, and NF reported throughout the paper are computed on 50K token vectors sampled uniformly from latent grids, using the 10M-token feature bank and $k = 64$ neighborhoods unless stated otherwise.

Intrinsic dimension (ID). We estimate intrinsic dimension with the two-nearest-neighbor estimator [Facco et al., 2017], which assumes locally uniform density and uses the ratio $\mu(u) = \|u - u_{(2)}\|/\|u - u_{(1)}\|$ between the second and first nearest-neighbor distances. The dataset-level estimate is the slope of the linear fit to $\{\log \mu_i,\ -\log(1 - F(\mu_i))\}_{i=1}^{N}$, where F is the empirical CDF. We report the rounded integer value over 50K sampled tokens.

Local linearity (LL). LL measures whether the local neighborhood is well approximated by a low-dimensional affine subspace. It is not computed with the adaptive 95% PCA dimension used to build the NormFuse projector, since that would make LL at least 0.95 by definition. Instead, we use a fixed diagnostic tangent budget $m = 32$ for all methods and compute

$$\mathrm{LL}(u) \;=\; \frac{\sum_{i=1}^{m} \lambda_i(u)}{\sum_{i=1}^{k-1} \lambda_i(u)} \;\in\; [0, 1], \qquad m = 32, \tag{15}$$

where $\{\lambda_i(u)\}_{i=1}^{k-1}$ are the ordered eigenvalues of the local neighborhood covariance. Since $m < k-1$ for the default $k = 64$, this diagnostic remains well defined even when the dataset-level ID estimate

exceeds the local neighborhood rank. $\mathrm{LL}(u) = 1$ indicates that the neighborhood variance is fully explained by this fixed tangent budget; lower values indicate curvature, noise, or locally heterogeneous directions beyond that budget. We report the mean over the 50K sampled tokens.

Normal fraction (NF). For a fused token with nonzero displacement $r_{ij} := z_{ij} - u_{ij}$ (including the fusion weight), NF measures how much of the residual magnitude is aligned with an independently re-estimated normal space of the base token manifold:

$$\mathrm{NF}(z_{ij}) \;=\; \frac{\|\widehat{\Pi}_N^{\mathrm{eval}}(z_{ij})\, r_{ij}\|}{\|r_{ij}\|} \;\in\; [0,1]. \tag{16}$$

Here $\widehat{\Pi}_N^{\mathrm{eval}}(z_{ij})$ is recomputed for evaluation by querying the feature bank around the fused token and running the same local PCA procedure; it is not the cached projector used to construct the NormFuse residual. Therefore the NormFuse NF is not algebraically forced to equal 1.0. Concat and Predictor-Add produce intermediate values depending on the tangent–normal composition of their residuals. NF is undefined for the Base-only setting ($r \equiv 0$), shown as "–" in all tables.

G.5 Decoder and Diffusion Backbone

The decoder maps the fused patch-token latent grid $z \in \mathbb{R}^{32\times32\times768}$ back to pixel space through (i) a point-wise linear projection from 768 to 256 channels, producing a $32 \times 32 \times 256$ spatial feature map, (ii) four upsampling blocks (each consisting of a nearest-neighbor $2\times$ upsample, a 3×3 convolution, GroupNorm, and SiLU), which progressively double the resolution from 32^2 to 512^2, and (iii) a final 1×1 convolution to 3 output channels. Skip connections from intermediate encoder layers are not used; the decoder operates solely on the fused patch-token grid.

The diffusion backbone is a Diffusion Transformer (DiT) [Peebles and Xie, 2023] trained with flow matching [Lipman et al., 2023]. For the fusion strategy comparison (Table 3), we use DiT-S/2 (12 layers, hidden dimension 384, 6 attention heads, 33M parameters) trained for 400K steps with batch size 256. For the scaled text-to-image experiments (Tables 1 and 2), we use a 4B-parameter DiT (32 layers, hidden dimension 2560, 32 attention heads, 8 KV heads, M-RoPE positional embeddings, and patch size 1). The matched RAE-Base model uses the same DiT, decoder family, text-conditioning stack, data mixture, and training schedule as NormFuse, but takes the final-layer DINOv3 token grid $h^{(12)}$ as the latent and removes the NormFuse residual branch. The matched Predictor-Add model uses the same scaled stack and multi-layer sources as NormFuse, but replaces the normal projector with unconstrained MLP residual addition. Text conditioning is provided by a frozen Qwen3-4B [Yang et al., 2025] language model whose final hidden states are projected through a learned linear layer and injected via cross-attention.

G.6 Training and Evaluation Protocol

Optimization. All training uses AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01) with a linear warmup over 5K steps to the peak learning rate (2×10^{-4} for the decoder, and 1×10^{-4} for the DiT), followed by cosine decay to 1×10^{-6}. Mixed precision (`bf16`) is used throughout.

Datasets and schedule. Ablation experiments are trained on COCO 2017 [Lin et al., 2014] at 512×512 with batch size 256. Scaled experiments are trained on $\sim$50M image–text pairs from BLIP-3o [Chen et al., 2025a] and DanQing [Shen et al., 2026], captioned at three granularities (short, medium, and long) in both Chinese and English (six captions per image), following a multi-stage progressive schedule: 256^2 (200K steps on the 50M image–text corpus), then 512^2 (150K steps), and finally 1024^2 (50K steps on a curated 1M-image high-quality subset).

Metrics. For the fusion strategy comparison, reconstruction metrics (PSNR and rFID) are computed on the COCO 2017 validation set (5K images) at 512×512. PSNR is averaged per image, and rFID is computed between the original validation images and their reconstructions. Generation quality (gFID) is computed on 50K images generated at 512×512 against COCO 2017 training statistics. Controlled ablation tables report compact mean values from three-seed averages; across the controlled seeds, the gFID standard deviation is below 0.15 for Table 3 and below 0.08 for the hyperparameter sweeps in Table 4. Small gFID margins are interpreted as directional evidence rather

than as standalone statistical claims. Large-scale GenEval and DPG-Bench tables include matched RAE-Base and Predictor-Add runs trained under the same recipe as NormFuse, and also quote representative reported scores from heterogeneous external systems for positioning. For the scaled experiments, we evaluate GenEval [Ghosh et al., 2023] (4 images per prompt) and DPG-Bench [Hu et al., 2024] (1 image per prompt) using 50 Euler sampling steps with classifier-free guidance scale 4.5.

Compute. All reconstruction autoencoder training and ablation experiments are conducted on 8 NVIDIA A100 80GB GPUs. The scaled text-to-image model is trained on 512 NVIDIA H20 GPUs for approximately 14 days across all stages (decoder pretraining, multi-resolution progressive training at 256^2, 512^2, and 1024^2, and aesthetic tuning). Geometric diagnostics (intrinsic dimension, local linearity, and normal fraction) require approximately 8 A100 hours, and feature bank construction and FAISS indexing require approximately 2 A100 hours.

H Relation to Prior Work

This appendix expands on Section 2 by placing NormFuse and the geometric framework in the context of three lines of work that are most closely related: manifold learning, score-based diffusion on manifolds, and representation autoencoder pipelines. Our goal is to make the boundary between what we inherit and what we contribute as explicit as possible.

H.1 Manifold Learning and Tangent–Normal Decompositions

Decompositions of local variation into tangent and normal components have a long history in nonlinear dimensionality reduction. Locally Linear Embedding [Roweis and Saul, 2000] and tangent-space alignment methods such as LTSA [Zhang and Zha, 2004] estimate local tangent spaces from nearest-neighbor sets and use them to parameterize a low-dimensional manifold from ambient samples. Isomap [Tenenbaum et al., 2000] and diffusion maps similarly exploit local geometric structure to recover global coordinates. Downstream, tangent propagation and related ideas in invariant feature learning encode the tangent–normal split as a training signal rather than as a post-hoc diagnostic.

What we share with this line of work is the basic geometric object, namely a locally linear approximation of the data manifold through the principal directions of a neighborhood patch. What we add is a purpose. The tangent–normal split is not used here to build a low-dimensional chart, nor to regularize a classifier, but to control the direction in which encoder residuals are added to each base patch token. In that sense our use of local PCA is tool-level, and the novelty lies in what the decomposition is applied to (token-wise residuals between encoder layers) and what it is applied for (preserving the base coordinate under tubular projection during fusion).

H.2 Score-Based Diffusion on Manifolds

A separate line of theory characterizes diffusion models when data are supported on a lower-dimensional submanifold. Riemannian score-based generative models [De Bortoli et al., 2022] extend score matching to general manifolds, and convergence analyses [De Bortoli, 2022, Chen et al., 2023] relate sample quality and the number of sampling steps to geometric properties of the data support, including its intrinsic dimension, curvature, and reach. More recent empirical work on diffusion generalization [Kamb and Ganguli, 2025, Aithal et al., 2024] attributes sample fidelity and memorization to local geometric regularity rather than to global density.

Proposition 3.2 lives in this tradition, and the qualitative dependence on intrinsic dimension, curvature, and reach is consistent with the kind of geometric factors that these works analyze in their own settings. What we add is not a tighter bound but an application of the underlying geometric intuition to a concrete design decision in latent diffusion. We use intrinsic dimension, curvature, and reach as diagnostic quantities that help anticipate which fusion strategies may hurt generation quality, and we turn that diagnostic use into an architectural choice through the Normal Projection Principle.

H.3 Representation Autoencoders and Frozen Visual Features

Representation Autoencoders [Zheng et al., 2025] replace the learned VAE encoder in a latent diffusion pipeline with a frozen visual foundation model, arguing that strong discriminative features form a navigable latent even when the encoder is never updated. SVG-T2I [Shi et al., 2025] extends this paradigm to text-to-image generation with DINOv3 features and introduces a predictor branch that adds detail residuals on top of a base layer. REPA [Yu et al., 2025] aligns the internal representations of a diffusion model with frozen VFM features during training, rather than replacing the latent space entirely. Latent Forcing [Baade et al., 2026] decouples coarse semantic structure from fine detail at different diffusion stages.

These works share with us the commitment to frozen VFM encoders and to fusing information across encoder depths. They differ from us in the principle that governs the fusion. SVG-T2I uses an unconstrained residual branch that permits both tangential and normal enrichment, whereas REPA and Latent Forcing operate at different levels of the generative stack and do not isolate a directional split. NormFuse is the first fusion rule in this family that explicitly enforces a normal projection. The 0.61 gFID gap between NormFuse and Predictor-Add in Table 3, together with the matched large-scale Predictor-Add comparison in Tables 1 and 2, supports the direction of enrichment as an important axis on which these pipelines differ.

I Qualitative Showcases

Figure 5 presents a gallery of images generated by NormFuse across a diverse set of text prompts, spanning multiple visual styles (photorealistic, cartoon, watercolor, oil painting, ink wash, and surrealist) and subject categories (human portraits, landscapes, still life, animals, food, and architecture). All samples are generated at 512×512 resolution using 50 Euler sampling steps with classifier-free guidance scale 4.5, matching the benchmark sampler settings, and the prompts are listed below for reproducibility. These examples illustrate that the geometry-guided fusion does not restrict the model to a particular visual domain. The base latent retains the broad semantic coverage of the DINOv3 encoder, while normal-space enrichment restores the fine-grained detail needed for faithful rendering across styles.

I.1 Prompts for Figure 5

The prompts used to generate Figure 5 are listed below in row-major order, from left to right and top to bottom:

1. *A cartoon-style boy in a blue T-shirt running joyfully along a sunlit countryside path with trees and a red-roofed house in the background.*
2. *Photorealistic portrait of a young Asian bride and groom gently embracing, soft studio lighting, neutral cream background, white lace wedding dress with veil.*
3. *Traditional Chinese watercolor painting of cherry blossoms draping over a stone arch bridge above a quiet stream.*
4. *A sun-drenched rice paddy field at sunset, golden grain heads in sharp focus, low-angle view with lens flare.*
5. *A serene turquoise mountain lake surrounded by forested green hills under a clear blue sky, mirror-like reflections.*
6. *A snow-capped mountain range mirrored perfectly on a still alpine lake at dawn, pastel sky, minimal color palette.*
7. *Surrealist oil painting of an inverted floating city suspended upside-down above a misty landscape, with tiny figures on the ground gazing upward.*
8. *A bustling fishing harbor at sunrise, fishermen sorting freshly caught silver fish in blue crates, seagulls circling overhead, moored boats in the background.*
9. *Traditional Chinese ink-wash landscape, layered distant mountains with a solo fisherman in a bamboo-hat punting a small boat on a tranquil river.*
10. *Classical oil painting of pink and white roses in a blue ceramic vase on a wooden table, soft warm lighting, painterly brushwork.*



Figure 5: **Qualitative showcases of NormFuse across diverse prompts and styles.** All corresponding prompts are listed in Appendix I.1.

11. *A tranquil lake at night under the Milky Way, stars reflected perfectly on the glassy water, distant silhouetted hills.*
12. *Close-up food photography of sliced Peking duck with crispy caramelized skin, accompanied by steamed pancakes and a small dish of hoisin sauce, on a white ceramic plate.*
13. *A golden retriever joyfully running through a park covered in autumn ginkgo leaves at golden hour, carrying a wooden stick in its mouth, soft backlight.*
14. *Oil-painting style scene of wooden rowboats on a calm canal, reflecting red-brick houses and arched stone bridges, a boatman punting at the stern.*
15. *A small rocky creek flowing between mossy stones, lush green trees along the banks, dappled daylight, cascading miniature waterfalls.*

J Broader Impacts and Asset Licenses

Broader impacts. NormFuse is a representation-based text-to-image model, so it inherits the standard positive and negative impacts of generative image systems. Potential positive uses include more controllable visual-content creation, research on representation autoencoders, and better understanding of how frozen visual features can be used in generative modeling. Potential negative uses include biased or stereotyped generations, copyright-sensitive imitation, misleading synthetic media, and other misuse common to text-to-image models. The method itself does not introduce a new modality of harm, but improved generation quality can amplify these risks. Practical deployments should therefore combine dataset filtering, prompt and output safety filters, provenance or watermarking mechanisms, and user-facing disclosure that images are synthetic.

Existing assets and licenses. We use COCO 2017 for controlled reconstruction and generation evaluation, ImageNet-1K for feature-bank construction and alignment-map training, BLIP-3o and DanQing image–text data for scaled training, DINOv3/DINOv2/SigLIPv2/MAE as frozen vision encoders, Qwen3-4B as a frozen text encoder, and GenEval and DPG-Bench for compositional evaluation. All assets are used under their respective research licenses or terms of use and are cited in the main text or appendix. COCO 2017 and ImageNet-1K are used for research evaluation and feature-bank construction under their public dataset terms; DINOv3, DINOv2, SigLIPv2, MAE, Qwen3-4B, BLIP-3o, DanQing, GenEval, and DPG-Bench are used under their corresponding model, dataset, and benchmark terms. We do not redistribute third-party data, model weights, or benchmark assets. Redistribution of trained checkpoints should follow the most restrictive license among the underlying datasets and pretrained models.

Compute and environmental cost. The controlled ablations are run on 8 A100 GPUs, while the scaled text-to-image model uses 512 H20 GPUs for approximately 14 days. We report these details in Appendix G to make the compute cost transparent. Future work should study whether the same geometric principle can reduce the amount of scaling required for competitive generation.

K Limitations

We group the current limitations into three categories, namely theoretical scope, implementation choices, and scaling.

Theoretical scope. Proposition 3.2 is a qualitative geometric bound. The constant C is not tight, and the manifold assumptions (uniform density bounds, sectional curvature bound, and positive reach) are idealizations of real latent distributions. Proposition 3.3 is a structural statement about tubular projection rather than a numerical bound on tangent-projected transport complexity. Strengthening either proposition into a sharper quantitative theorem, for example by matching the bound to concrete score estimation error or by quantifying how normal enrichment changes the induced metric, would tighten the link between theory and practice.

Manifold-assumption gap. The propositions assume a smooth compact submanifold with bounded curvature and positive reach, whereas real DINOv3 token latents are only approximately manifold-valued: density varies across semantic clusters and spatial contexts, the reach is not uniform, and

the local PCA tangent dimension used by the projector is heterogeneous (we estimate it per token through the 95%-variance threshold, not as a single dataset-wide d). The empirical diagnostics (ID, LL, NF) should therefore be read as coarse geometric summaries rather than exact instantiations of the theory, and the ordering they produce across fusion strategies is what we rely on, not their absolute values. Closing this gap, either by measuring deviations from the manifold model directly or by extending the analysis to stratified or density-varying supports, is a natural next step.

Implementation choices. Three aspects of the current NormFuse implementation are approximations whose impact is mitigated but not eliminated. (i) Local PCA estimates the tangent space from a k-nearest-neighbor set and can degrade in regions of high curvature or sparse data; incorporating curvature-aware estimators, such as local quadratic fits, may improve accuracy. (ii) We choose the tangent dimension by a 95% variance threshold, which is well motivated but globally fixed; adaptive criteria based on eigenvalue gaps could better handle heterogeneous regions. (iii) The main fusion uses a single scalar α shared across tokens. The exploratory confidence-adaptive variant above has a small observed gain, but does not establish statistical significance or directly estimate reach.

Scaling and generality. The controlled cross-encoder comparison in Appendix D covers DINOv2, SigLIPv2, and MAE. The scaled evaluation includes matched DINOv3 models with 50M pairs and SigLIPv2 models at the same 4B capacity with 5M pairs. A second-encoder 50M replication and scaled Concat comparison remain open. The empirical evaluation is limited to text-to-image generation; video and content-preserving editing remain future work.

L Future Work

Beyond the text-to-image setting studied in the main paper, we highlight three concrete directions for future work.

Other modalities and generative stacks. The geometric quantities driving our analysis, namely intrinsic dimension, curvature, and reach, are properties of the latent manifold rather than of any specific modality or sampler. In video generation, residuals between encoder layers can be decomposed into components that preserve the per-frame base coordinate and components that shift it, and the Normal Projection Principle extends by penalizing the latter. The same argument applies to other generative stacks such as standard Gaussian diffusion, consistency models, and rectified flow, since Proposition 3.3 is purely a statement about tubular projection.

Normal projection as a soft training signal. NormFuse enforces the Normal Projection Principle through a hard projection at fusion time. A softer alternative is to use the normal fraction NF as a training signal, encouraging the residual branch to produce high-NF residuals through an auxiliary loss rather than projecting at every forward pass. A soft penalty defined with the same projector would still require tangent estimation and would not itself remove retrieval or PCA preprocessing; reducing that cost would require an additional learned approximation or distillation step.

Controllable editing via normal enrichment. Because normal residuals preserve the base coordinate under tubular projection, they add detail without changing the identity of the underlying base feature. This suggests an editing protocol in which the base latent is held fixed to preserve coarse content while only the normal residual is varied to modulate detail or style, for instance by interpolating between two aligned detail sources while keeping the base feature constant.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction make two scoped claims: (i) latent quality can be interpreted through the navigability–invertibility trade-off, supported by the geometric framework and synthetic sanity checks; and (ii) NormFuse instantiates this view as a normal-projected multi-layer RAE latent, supported by controlled fusion ablations and matched large-scale RAE-Base and Predictor-Add comparisons.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations appendix discusses the qualitative nature of the theoretical bounds, the gap between smooth-manifold assumptions and real token latents, implementation choices such as local PCA, the variance threshold, k-NN size and fusion weight, and the remaining scaling/generalization limits.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each proposition states its assumptions explicitly, including smooth compact submanifolds, bounded curvature, positive reach, density regularity, and valid tubular neighborhoods. The appendices provide the proof of the tangent-projected transport-complexity bound and the tubular-projection proof for the normal-versus-tangential enrichment statement, together with synthetic verification.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the NormFuse architecture, including base/detail layers, alignment maps, FAISS feature bank, k-NN PCA, normal projection, and fusion weight α. Section 5.1 and the implementation-details appendix report datasets, model configurations, optimizer, learning-rate schedule, batch size, precision, progressive training stages, sampling settings, and evaluation protocols.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and pretrained weights are not released at submission due to ongoing responsible-release review. The architecture, training protocol, hyperparameters, and evaluation setup are described in Section 4, Section 5.1, and the implementation-details appendix; we plan to release code and configs after the review is complete.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 5.1 and the implementation-details appendix specify data splits, controlled and scaled backbones (DiT-S/2 and 4B DiT), AdamW settings, warmup and cosine decay, batch size, bf16 precision, progressive $256^2/512^2/1024^2$ training, sampling steps, guidance scale, and evaluation benchmarks. Choices of α and k are discussed in Section 5.3, and the PCA variance threshold is evaluated in Appendix G.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Controlled ablations are averaged over three seeds with the encoder, decoder, backbone, training, sampler, and evaluation held fixed; the across-seed gFID standard deviation is below 0.15, smaller than the 0.61 gap between NormFuse and Predictor-Add. Large-scale GenEval and DPG-Bench results include matched NormFuse/RAE-Base/Predictor-Add comparisons under the same evaluation protocol, while scores for heterogeneous external systems are quoted for positioning rather than treated as statistically controlled comparisons.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The implementation-details appendix reports the main compute budget: autoencoder training and controlled ablations on 8 NVIDIA A100 80GB GPUs, the scaled text-to-image model on 512 NVIDIA H20 GPUs for approximately 14 days, geometric diagnostics requiring approximately 8 A100-hours, and FAISS feature-bank construction requiring approximately 2 A100-hours.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics. The experiments use released datasets, benchmarks, and pretrained checkpoints under their stated research-use terms. The work does not involve human-subject experiments, crowdsourcing, or collection of personally identifiable information.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impacts appendix discusses positive uses such as controllable visual-content creation and research on representation-based generative models, as well as negative risks common to text-to-image systems, including biased or stereotyped generations, copyright-sensitive imitation, and misleading synthetic media. It also notes practical mitigations such as dataset filtering, output safety filters, provenance or watermarking mechanisms, and user-facing disclosure.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No trained model, pretrained checkpoint, or new dataset is released with this submission. Any future release will be subject to responsible-release review and may include gated access, usage guidelines, safety filters, and watermarking or provenance mechanisms as appropriate.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the external assets used in the experiments, including DINOv3, Qwen3-4B, COCO 2017, ImageNet, BLIP-3o, DanQing, GenEval, DPG-Bench, and FAISS, and states that they are used under their respective research licenses or terms of use. Asset-license information is summarized in the appendix.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission does not release new datasets, pretrained checkpoints, or standalone assets. The paper introduces a method and reports experiments using existing datasets, benchmarks, and pretrained models.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments, human-subject studies, user studies, or human annotation tasks.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subject research is involved, so IRB approval or an equivalent ethics review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The scaled text-to-image system uses Qwen3-4B as a frozen text encoder via cross-attention, and this is disclosed in Section 5.1 and the implementation-details appendix. The paper does not rely on an LLM to generate the proposed method, theoretical analysis, experimental conclusions, or reported results.