STREAM: Stochastic Riemannian Flow Matching with Anisotropic Decoder for Digital Histopathology Image Generation

Won June Cho **Daeky Jeong** **Hyeongyeol Lim** **Hongjun Yoon**
DEEPNOID Inc.
{wjcho, dkjeong, hylim, hyoon}@deepnoid.com

Abstract

Synthetic histopathology image generation addresses patient-privacy concerns and the growing data demands of foundation models. Existing *state-of-the-art* histopathology generative models use pretrained Vision Foundation Models (VFMs) as conditioning signals. We show this yields *conditioning-dominated diversity*: on TCGA-BRCA, 62–75% of their output diversity is attributable to the conditioning signal rather than the learned latent space, while *de novo* synthesis still requires a VFM embedding at inference. We instead use histopathology VFMs as the latent space itself: their patch tokens are ℓ_2-normalized on the unit hypersphere $\mathcal{S}^{d-1}$ with strong angular dominance and intrinsic curvature, motivating a Riemannian formulation. We present *STREAM*, the first framework to apply Riemannian flow matching in the histopathology domain, in two stages: 1) a bridge-type stochastic perturbation that establishes per-token rectifiability on $\mathcal{S}^{d-1}$ for training a Diffusion Transformer, and 2) a novel decoder training design whose noise covariance is anisotropic in the left-singular basis of the per-token tangent-projected velocity-field Jacobian, spending a large robustness budget on its low-response directions and a small one on its high-response directions. Across TCGA-BRCA, TCGA-COADREAD and SPIDER-skin, STREAM achieves *state-of-the-art* gFID and ranks first on nearly all histopathology-specific metrics as well. Code and a public gallery of generated images are available at https://chokevin8.github.io/STREAM-Patho/.

1 Introduction

Synthetic medical image generation addresses critical challenges in computational histopathology: patient privacy concerns that limit data sharing across institutions [62] and the growing need for training data as large Vision Foundation Models (VFMs) and Vision-Language Models are applied to histopathology [13, 42, 66, 68, 78].

For high-resolution image generation, latent diffusion models [23, 48, 51] have dominated the space. A recent line of work redesigns the underlying tokenizer [12, 35, 71, 77], as the choice of latent space is increasingly recognized as critical to downstream generation quality. The underlying observation across these works is that latents generated via Variational Autoencoders (VAEs) primarily emphasize low-level details [11], with linear classifiers on VAE features substantially underperforming those on representation-encoder features [12, 77]. We confirm this pattern transfers to the histopathology setting: linear probing on downstream binary classification (Table 2) shows that the VAE encoders used by current state-of-the-art histopathology generative models ZoomLDM [74] and PixCell [73] yield substantially lower performance than histopathology VFMs [13, 68, 78]. Both methods condition diffusion on the [CLS]-token of a pretrained histopathology VFM, which we show empirically (Section 3.1) leads to conditioning-dominated diversity: most of the output diver-

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

sity is attributable to the conditioning signal rather than to the learned latent space, which therefore carries little semantic content, and any *de novo* synthesis requires a VFM embedding at inference, which is impractical for clinical deployment. Moreover, digital histopathology particularly benefits from unconditional generative approaches due to the high cost of expert pathologist annotation [8] and the combinatorial diversity of tissue morphology [38]. Therefore, using histopathology VFM patch-token features directly as the generative latent space resolves both issues, but requires accounting for their geometry: the ℓ_2-normalized features lie on the unit hypersphere $\mathcal{S}^{d-1}$ [10, 31], with strong angular dominance (Section 3.1) and substantial intrinsic curvature (Section 3.2), motivating Riemannian flow matching (RFM). STREAM is the first framework to apply RFM with VFM encoders in the histopathology domain. Our main contributions are:

1. **Motivating RFM in the histopathology domain** (Section 3): we quantify conditioning-dominated diversity in current state-of-the-art histopathology models and provide empirical evidence on the spherical geometry of histopathology VFM features that justifies a Riemannian formulation.
2. **Bridge-type stochastic RFM** (Section 4.2): a tangent-Gaussian perturbation of the spherical linear interpolation (SLERP) geodesic on $\mathcal{S}^{d-1}$ with a Brownian-bridge schedule that vanishes at the endpoints, providing per-token rectifiability of the bridge construction at the population level (Theorem 1).
3. **Anisotropic decoder training** (Section 4.3): a novel decoder training design whose noise covariance is shaped by the singular value decomposition (SVD) of the velocity-field Jacobian of the trained Diffusion Transformer (DiT) [48]: small noise along the high-energy left singular directions U_H, which span the directions of largest velocity-field response, to preserve reconstruction fidelity, and large noise along the complementary low-response directions U_L.

2 Related Work

Diffusion models for histopathology generation. Diffusion-based generative modeling for histopathology spans large-image and multi-scale synthesis [20, 32, 73]. Other works incorporate text or mask conditioning for guided synthesis like pathology reports [72], cell topology [69], and joint nuclei image–label co-synthesis [45]. PathDiff [5] conditions on unpaired text and mask; like ZoomLDM [74] and PixCell [73], it is a conditional VAE-latent method. At inference it accepts text, a mask, or both, and provides no unconditional branch, so it cannot sample without a conditioning signal. While digital histopathology routinely leverages pretrained VFMs for downstream tasks (classification, segmentation, retrieval), no current state-of-the-art histopathology generation model exploits these representations as the generative latent space. Therefore, STREAM intends to focus on purely unconditional histopathology generation: a diffusion model that faithfully learns the histopathology data distribution from a VFM-encoded latent space without external conditioning.

Latent space modification in diffusion model training. VFMs have emerged as feature extractors for generative models via two paradigms. Paradigm 1 uses VFM features as training losses: REPA [75] aligns DiT features with DINOv2 [47], REPA-E [35] extends to end-to-end VAE tuning, and others similarly leverage VFM representations [53, 63, 71]. Paradigm 2 uses VFM features as the latent space: RAE [77] trains DiTs directly on representation-encoder features with noise-augmented decoders, SVG [52] similarly uses frozen DINOv3 [54] features, concatenated with a trained residual branch, and related works explore fine-tuning representation encoders under a semantic-preservation loss [11, 56]. Decoder regularization via masking or denoising losses has been shown beneficial for jointly-trained tokenizers [30, 50, 70]. We adopt Paradigm 2 because histopathology-specific VFMs [13, 68, 78] produce features naturally suited to Riemannian geometry once ℓ_2-normalized. STREAM differs from approaches that jointly adapt the tokenizer encoder–decoder to improve reconstruction fidelity [11, 35] by keeping the histopathology VFM frozen and instead recovering diffusion-friendliness via geometry: a stochastic bridge on $\mathcal{S}^{d-1}$ paired with anisotropic decoder regularization (Section 4.3).

Flow matching and Riemannian extensions. Flow matching (FM) [1, 37, 40] provides a simulation-free alternative for generative modeling, scaled to high-resolution synthesis [18, 43]. Rectified flow [39, 40] reframes generation as straight transport between marginals. On manifolds, the canonical formulation is the SLERP conditional flow on $\mathcal{S}^{d-1}$ [14], inherited by subsequent RFM work. In high-dimensional image generation, RJF [31] runs on frozen-VFM latents

Table 1: Conditioning-dominated diversity on BRCA (Eq. (1)).

Model	$H(X\|C)$	$H(X)$	ρ_{cond}	CVS	SPEC-diff
ZoomLDM	1.28	5.17	75%	3.59	0.067
PixCell	1.81	4.78	62%	6.08	0.102

Table 2: Linear probing on SPIDER-breast (benign vs. malignant). AUROC is the area under the receiver operating characteristic curve.

Model	d	AUROC	Acc_{dir}	Acc_{mag}	f_{angular}
UNI	1024	0.995	0.993	0.583	0.999
UNI2-h	1536	**0.997**	**0.995**	0.603	0.973
DINOv3-L	1024	0.937	0.930	0.514	0.999
PixCell VAE	16	0.640	0.832	0.549	0.994
ZoomLDM VAE	3	0.624	0.609	0.543	0.813

and reweights the velocity-matching loss by a Jacobi-field factor; HAE [10] replaces Vision Transformer (ViT) patchification with a trained convolutional stem and aligns to the VFM by a cosine loss; and SFM [34] renormalizes autoencoder latents onto the sphere. Each defines the conditional path as a deterministic geodesic between prior and data and regresses the velocity given in closed form by the time-derivative of that path, so the target lies in the tangent space by construction. STREAM's design differs further in three aspects: (i) perturbing the SLERP path with a stochastic bridge and regressing a normalized endpoint under the chord loss (Eq. (5)), (ii) exploiting the Riemannian formulation for anisotropic decoder training, and (iii) jointly addressing rectifiability and decoder–generator interaction in the RFM setting. Direct numerical comparison with these RFM works is not possible, since none of them has been open-sourced for reproducible training to date.

3 Why RFM for Histopathology?

3.1 Conditioning-Dominated Diversity and Angular Dominance in VFM Feature Spaces

State-of-the-art histopathology generative models ZoomLDM [74] and PixCell [73] condition diffusion models on VFM embeddings $c = E_{\text{VFM}}(I)$. We formalize a conditioning dominance ratio via the entropy decomposition $H(X) = H(X \mid C) + I(X;C)$:

$$\rho_{\text{cond}} = \frac{I(X;C)}{H(X)} = 1 - \frac{H(X \mid C)}{H(X)}, \tag{1}$$

estimated via the Vendi Score [19], the Conditional Vendi Score (CVS) [24], and SPEC-diff [25]. The Vendi Score and CVS exponentiate the Shannon entropy of the normalized eigenvalues of a similarity kernel matrix on UNI embeddings. The log Vendi Score estimates the total output diversity $H(X)$, and the log CVS estimates the diversity left at fixed conditioning, $H(X \mid C)$. Their difference $I(X;C)$ is the amount due to conditioning, and ρ_{cond} is its share of $H(X)$. Both models exhibit conditioning-dominated diversity (Table 1): 62–75% of output diversity is attributable to the conditioning signal rather than the learned latent space. This is compounded by VAE-extracted latents being weakly structured semantically [11, 12, 77]: linear-probe AUROC on SPIDER-breast [46] for malignant vs. benign tissue is ≥ 0.937 for VFMs vs. ≤ 0.640 for VAEs (Table 2; full protocol in Appendix A.3). This motivates using VFM features directly as the generative latent space.

We next ask where in VFM features the semantic information resides via 1-nearest-neighbor (1-NN) angular vs. radial perturbation on SPIDER-breast (full protocol in Appendix A.4). For both histopathology VFMs and the natural-image-domain DINOv3-L, $f_{\text{angular}} > 0.97$ (Table 2): class-discriminative information lives in the angular structure of $\mathcal{S}^{d-1}$, directly motivating RFM.

3.2 Intrinsic Geometry of VFM Feature Manifolds

Following Xiong et al. [67], we analyze the intrinsic geometry of histopathology VFM features on TCGA-BRCA (hereafter BRCA) for UNI (d=1024) and UNI2-h (d=1536). Spectral analysis yields a high covariance effective rank ($R_{\text{eff}} = 252$ for UNI, 265 for UNI2-h), and tangent drift exceeds 0.72 at hop 1, direct evidence of manifold curvature (Appendix A.5). Drift this large means straight-line interpolants leave the manifold, and we ℓ_2-normalize the features onto $\mathcal{S}^{d-1}$, so RFM on the hypersphere is the better-matched formulation. We use UNI (d=1024) henceforth: UNI2-h buys 5% more effective rank for 50% more embedding dimension.

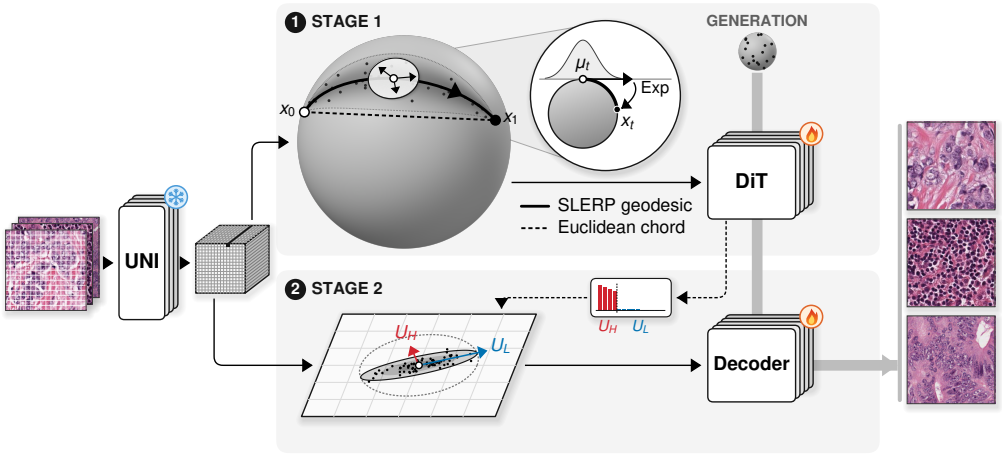


Figure 1: **Overview of STREAM.** A frozen histopathology VFM encoder maps each input patch to N=256 tokens on the unit hypersphere $\mathcal{S}^{d-1}$; a DiT learns to transport a uniform source distribution on $(\mathcal{S}^{d-1})^N$ to the data distribution along bridge-perturbed geodesics. A separately-trained anisotropic decoder reconstructs histopathology images from the generated features.

4 Method: STREAM

FM [37] learns a time-dependent velocity field $v_\theta(x,t)$ generating a probability path p_t from a source p_0 to data p_1. The intractable marginal FM loss shares its gradient with the tractable conditional FM (CFM) loss $\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,z,X_t|z}[\|v_\theta - u_t\|^2]$, where $u_t(x \mid z)$ is the conditional velocity along a known path. RFM is the same construction on a Riemannian manifold, with each Euclidean vector-space operation replaced by its intrinsic analogue. On the unit hypersphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, the Riemannian counterparts of vector difference, displacement and distance are the logarithmic map Log_x, the exponential map Exp_x and the geodesic distance d_g. The conditional path becomes a geodesic [14], SLERP with constant speed $\|\dot{\mu}_t\|_g = \Omega$:

$$\text{SLERP}(x_0, x_1, t) = \frac{\sin((1-t)\Omega)}{\sin\Omega}\, x_0 + \frac{\sin(t\Omega)}{\sin\Omega}\, x_1, \quad \Omega = d_g(x_0, x_1), \tag{2}$$

with conditional velocity $u_t(x \mid x_1) = \text{Log}_x(x_1)/(1-t)$. Background on optimal transport (OT), the curvature-dimension condition, and rectifiability theory is in Appendix B.

4.1 Problem Setup

Our pipeline (Fig. 1) adopts the RAE framework [77]:

1. **Encoder** $E : \mathcal{X} \to (\mathcal{S}^{d-1})^N$: A frozen VFM (UNI [13], a ViT-L/16 [16]) mapping 256×256 histopathology patches to $N = (256/16)^2 = 256$ ℓ_2-normalized patch tokens of dimension $d = 1024$.
2. **Generator** G_θ: A DiT learning to generate features $\hat{\mathbf{Z}} \sim p_\theta$ approximating $p_{\text{data}} = E_\#(p_{\text{real}})$.
3. **Decoder** $D_\phi : (\mathcal{S}^{d-1})^N \to \mathcal{X}$: A ViT-based decoder reconstructing images from features.

The three components interact as follows. The encoder E produces clean VFM features $\mathbf{Z}_1 = E(I) \in (\mathcal{S}^{d-1})^N$ for each training image I. The generator G_θ is trained to recover $\mathbf{Z}_1$ from a bridge-perturbed interpolant on $(\mathcal{S}^{d-1})^N$: a uniform source $\mathbf{Z}_0 \sim p_0$ is interpolated toward $\mathbf{Z}_1$ along the SLERP geodesic to produce μ_t, then perturbed by tangent-Gaussian noise $X_t = \text{Exp}_{\mu_t}(\sigma(t)\epsilon)$ with the bridge schedule $\sigma(t) = \sigma_{\max}\sin(\pi t)$ peaking at t=0.5 and vanishing at the endpoints (Section 4.2). The decoder D_ϕ is trained separately, after G_θ, on clean VFM features perturbed by anisotropic noise whose covariance is shaped by the SVD of the trained DiT's velocity-field Jacobian (Section 4.3). This couples the decoder's directional robustness budget to the DiT velocity field's left-singular directions of largest and smallest response. At generation time, $\mathbf{Z}_0 \sim p_0$ is integrated through G_θ via Riemannian midpoint steps on $(\mathcal{S}^{d-1})^N$ to produce $\hat{\mathbf{Z}}_1 \sim p_\theta$, and $D_\phi(\hat{\mathbf{Z}}_1)$ outputs the image. Both perturbations act at training time only: at generation time D_ϕ reads $\hat{\mathbf{Z}}_1$ unperturbed.

The generative model operates on the product manifold $(\mathcal{S}^{d-1})^N$ with the product metric $g = \bigoplus_{n=1}^{N} g_n$, total dimension $N(d-1) = 261{,}888$. The source distribution is $p_0 = \text{Uniform}((\mathcal{S}^{d-1})^N)$, the unique stationary measure of Brownian motion on $(\mathcal{S}^{d-1})^N$, geometrically canonical by the Bakry–Émery curvature-dimension condition $\text{CD}(d-2, \infty)$ [2]. Geodesics, exponential maps, and all constructions decompose as per-token operations on the product manifold; we present the theory for a single copy of $\mathcal{S}^{d-1}$ throughout.

4.2 Stochastic Bridge RFM

RFM [14] on $\mathcal{S}^{d-1}$ trains a velocity field $v_\theta(x_t, t)$ via the CFM loss along deterministic SLERP paths,

$$\mathcal{L}_{\text{RFM}}(\theta) = \mathbb{E}_{t\sim U[0,1],\,(x_0,x_1)\sim\pi}\left[\left\|v_\theta(x_t,t) - u_t(x_t \mid x_1)\right\|_g^2\right], \tag{3}$$

with $x_t = \text{SLERP}(x_0, x_1, t)$ and $u_t(x_t \mid x_1) = \text{Log}_{x_t}(x_1)/(1-t)$. Two limitations of Eq. (3) motivate our bridge construction: (L1) no general rectifiability guarantee [21], addressed by Theorem 1; and (L2) the disconnected-intermediate-support obstruction [21, Prop. 10] that arises when data clusters on $\mathcal{S}^{d-1}$ and is not removed by a uniform source (Appendix B.5), resolved by Corollary 1.

Definition 1 (Stochastic bridge conditional path)**.** *Given the independent coupling $\pi = p_0 \otimes p_1$ on $\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$, $(x_0, x_1) \sim \pi$, and noise schedule $\sigma(t) = \sigma_{\max}\sin(\pi t)$:*

$$\mu_t = \text{SLERP}(x_0, x_1, t), \qquad x_t = \text{Exp}_{\mu_t}\big(\sigma(t)\,\epsilon\big), \quad \epsilon \sim \mathcal{N}(0, \Pi_{\mu_t}), \tag{4}$$

where $\Pi_{\mu_t} = I - \mu_t\mu_t^\top$ is the tangent-space projector. Throughout, t=0 corresponds to the source distribution p_0 (uniform noise) and t=1 to the data distribution p_1, following the FM convention [37, 39].

The stochastic noise schedule $\sigma(t) = \sigma_{\max}\sin(\pi t)$ is the $n = 1$ Karhunen–Loève (KL) mode of the Brownian bridge (the unique sinusoid strictly positive on $(0,1)$ with vanishing endpoints), with $\sigma_{\max} \lesssim \pi/\sqrt{d-1}$ ensuring perturbations remain inside the injectivity radius of $\mathcal{S}^{d-1}$. Vanishing endpoints $\sigma(0) = \sigma(1) = 0$ are critical: the bridge must not perturb the source (t=0) or data (t=1) endpoints, so $X_0 = x_0$ and $X_1 = x_1$ exactly and the network sees clean source and data without conflating them with noise (Appendix C.1).

Training loss and pipeline. We train with a chordal (extrinsic Euclidean) loss, hereafter the chord loss, replacing the tangent-space geodesic distance with the ambient ℓ_2 distance between normalized predictions, using x-prediction [36] with normalized output $\hat{x}_1 = f_\theta(x_t, t)/\|f_\theta(x_t, t)\|$:

$$\mathcal{L}_{\text{chord}}(\theta) = \mathbb{E}_{t,\,(x_0,x_1)\sim\pi,\,\epsilon}\left[\frac{\|\hat{x}_1 - x_1\|^2}{(1-t)^2}\right]. \tag{5}$$

Sample by sample, $\mathcal{L}_{\text{chord}}$ and the natural Riemannian (geodesic) loss have the same minimizer $\hat{x}_1 = x_1$ and the same inference velocity $v_\theta(x_t, t) = \text{Log}_{x_t}(\hat{x}_1)/(1-t)$, but $\mathcal{L}_{\text{chord}}$ avoids an arccos-induced singularity near $t \to 1$ (Appendix C.2). Concretely, f_θ is a LightningDiT-XL [71] trained to predict per-token data endpoints on $\mathcal{S}^{d-1}$ from bridge-perturbed VFM embeddings (Algorithm 1). At inference, generation uses 25 midpoint Euler steps on $(\mathcal{S}^{d-1})^N$ (Algorithm 3), followed by anisotropic decoding (Section 4.3). The bridge smooths the endpoint regression with bandwidth $\sigma(t)$ rather than defining a path we integrate, so the marginal field it induces differs from the inference field (Appendix B.4).

Proposition 1 (Full support)**.** *Under the stochastic bridge (Definition 1) with $\sigma_{\max} > 0$, and for any coupling π giving no mass to the antipodal locus $\{(x_0, x_1) : x_1 = -x_0\}$, $\text{law}(X_t)$ has full support on $\mathcal{S}^{d-1}$ for all $t \in (0,1)$.*

Theorem 1 (Bridge perturbation: three consequences)**.** *Under the hypotheses of Proposition 1, on a single $\mathcal{S}^{d-1}$ the bridge yields three separate properties:*

- *(i)* ***Full support*** *(Proposition 1):* $\text{supp}(\text{law}(X_t)) = \mathcal{S}^{d-1}$ *for all $t \in (0,1)$.*
- *(ii)* ***Smooth conditional densities****: the conditional distribution of (X_0, X_1) given $X_t = x$ admits a smooth density for $\text{law}(X_t)$-a.e. x, the step from which part (iii) obtains rectifiability in the sense of Hertrich et al. [21, Def. 1].*

(iii) ***Velocity-field regularity****: for the integration cutoff $\epsilon_t > 0$, the marginal velocity $v_t(x)$ is smooth on $\mathcal{S}^{d-1} \times [\epsilon_t, 1-\epsilon_t]$, and $\dot{x} = v_t(x)$ admits a unique solution there for any initial condition.*

All three lift unchanged from $\mathcal{S}^{d-1}$ to the product $(\mathcal{S}^{d-1})^N$, with N entering only the constants.

Full proof is in Appendix B.3. No region of $\mathcal{S}^{d-1}$ is unreachable at an interior time, which removes the disconnected-support obstruction (L2) that deterministic SLERP leaves in place. Lemma 2 guarantees the density that part (ii) conditions on is positive almost everywhere, and the bridge kernel on $\mathcal{S}^{d-1}$ supplies the smooth conditional densities themselves. Part (ii) together with the bounded-velocity bound then gives part (iii) by the Picard–Lindelöf theorem, the unique solution of $\dot{x} = v_t(x)$ that Hertrich et al. [21, Def. 1] call rectifiability. The bridge removes the binding obstruction that Hertrich et al. [21, Prop. 10] identify by supplying the intermediate-marginal property of part (i). The lift to $(\mathcal{S}^{d-1})^N$ carries no cross-token assumption, because the SLERP geodesic, the exponential map and the tangent Gaussian act per token.

4.3 Spectral-Informed Anisotropic Decoder Regularization

Without noise injection, a decoder trained on clean VFM features must handle generated features $\hat{z} \sim p_\theta$ that may fall in regions of $\mathcal{S}^{d-1}$ underrepresented in the training distribution. Isotropic noise augmentation improves generation quality, measured by the Fréchet Inception Distance (FID) of generated images (gFID), but degrades reconstruction quality, measured by the FID of reconstructed images (rFID) [77]; with covariance $\sigma^2 I$, that cost falls on all directions equally. Empirically (Table 4), allocating noise according to the SVD of the trained DiT's velocity-field Jacobian $J(z) = \nabla_z v_\theta(z, t)$ yields complementary rFID/gFID gains; we interpret this as the SVD decomposing each token's tangent space into high-energy directions U_H, the top-k^* left singular directions by spectral energy, and low-energy directions U_L, the rest. Because v_θ is a tangent field, the cached J is a single token's tangent-projected diagonal block, an endomorphism of $T_z\mathcal{S}^{d-1}$; its left factor is an orthonormal basis of $\mathrm{range}(J)$, so U_H and U_L split the velocity field's response directions rather than its input-perturbation sensitivities (Appendix G.3). We use this partition as an empirical budget for the decoder's directional robustness and inject asymmetric noise:

$$\Sigma_{\text{noise}}(z) = \sigma_H^2 \, U_H U_H^\top + \sigma_L^2 \, U_L U_L^\top, \quad \sigma_H \ll \sigma_L, \tag{6}$$

where U_H holds the top-k^* singular vectors of $J(z)$ and U_L their tangent complement, with k^* defined by the energy threshold τ: $k^* = \min\{k : \sum_{i \le k} s_i^2 / \sum_i s_i^2 \ge \tau\}$ where s_i is the median cached spectrum. The basis $\{U_H(\bar{z}_c), U_L(\bar{z}_c)\}$ is precomputed once from the trained DiT at the single time anchor $t = 0.99$: training features are k-means clustered into K centroids $\bar{z}_c$, the per-(centroid, token) SVD of $J(\bar{z}_c, t{=}0.99)$ is computed via batched forward-mode automatic differentiation (AD) with m random tangent probes, and at training time each sample is assigned to its nearest centroid and receives the corresponding basis $U_H(\bar{z}_c)$. Crucially, no discrete spectral gap is required; only the continuous energy ordering matters. Full computation and deployed values are in Appendix D.

4.4 Decoder Training

The decoder $D_\phi : (\mathcal{S}^{d-1})^N \to \mathbb{R}^{H \times W \times 3}$ is a ViT-XL mapping VFM embeddings on the product hypersphere back to histopathology images, trained from scratch with anisotropic noise injection from the start (Algorithm 2), using a 4-component loss with a Learned Perceptual Image Patch Similarity (LPIPS) term:

$$\mathcal{L}_{\text{dec}} = \lambda_{\ell_1} \|\hat{I} - I\|_1 + \lambda_{\text{LPIPS}} \text{LPIPS}(\hat{I}, I) + \lambda_{\cos}(1 - \langle z_{\text{rt}}, z \rangle) + \omega_G \lambda_{\text{adapt}} \mathcal{L}_{\text{adv}}, \tag{7}$$

where $\hat{I} = D_\phi(\tilde{z})$ is the reconstruction from noisy features $\tilde{z} = (z + n)/\|z + n\|$ with $n = \Pi_z \eta$, $\eta \sim \mathcal{N}(0, \Sigma_{\text{noise}}(\bar{z}_c) + \sigma_L^2 \bar{z}_c \bar{z}_c^\top)$, $z_{\text{rt}} = E(D_\phi(\tilde{z}))$ is the round-trip re-encoding, and $\mathcal{L}_{\text{adv}}$ is an adversarial loss against a Kaiko ViT-S/8 discriminator [29], with adaptive weight following VQ-GAN [17]. The cosine round-trip loss targets the clean features z (not $\tilde{z}$), forcing the decoder to denoise. Training follows a standard VQGAN-style staged schedule [17, 51], with anisotropic noise from the beginning so the decoder jointly allocates capacity for reconstruction and noise robustness. Deriving the noise covariance from the SVD of the velocity-field Jacobian (Eq. (6)) creates a deliberate coupling: the spectral decomposition governing the DiT's directional response determines the decoder's robustness allocation (Table 4). Training the same decoder on the right singular vectors instead is worse on both metrics (Appendix G.3), so STREAM keeps the left factor.

5 Experiments

5.1 Implementation Details

Datasets and encoder. We use 12.2M patches from BRCA [58], following the patch set of ZoomLDM [74], and 3.4M patches from TCGA-COADREAD [57] (hereafter COADREAD), both at 20× magnification (256×256 pixels). We extract all training VFM features with UNI [13] (ViT-L/16, $d = 1024$, $N = 256$ tokens). We train and evaluate on another independent, non-TCGA dataset at 20× magnification, SPIDER-skin [46], and report that comparison in Appendix E.1.

Training. We train sequentially in two stages on 8× NVIDIA H200 GPUs. In Stage 1, we train the LightningDiT-XL [71] (∼676M parameters) with the extracted UNI patch embeddings for 95k steps with batch size 1024 using the stochastic bridge FM loss (Algorithm 1). In Stage 2, we extract the cached SVD of the velocity-field Jacobian from the trained DiT first, and use it to train a ViT-XL decoder [16, 77] from scratch with anisotropic noise for 90k steps with a batch size of 512. Full details are in Appendix D.

Baselines. We benchmark against the state-of-the-art histopathology-domain models ZoomLDM [74] and PixCell [73], and against three methods that build diffusion models on VFM-encoded latents: RAE [77] as our Euclidean baseline, SVG [52], which augments RAE's design with a learned residual encoder on top of the frozen VFM features, and REPA-E [35], which uses the VFM as a representation-alignment target that tunes a VAE end-to-end. We train all baselines on BRCA and COADREAD with the default hyperparameters and settings of their respective papers. Every method in Table 3 uses the same UNI encoder, so the comparison is encoder-controlled. RAE, SVG and REPA-E are therefore retrained here on UNI (d=1024) rather than on their native natural-image encoders. UNI enters each method in a different role: as the generative latent for STREAM, RAE and SVG, with no VAE; as a conditioning input on a VAE latent for ZoomLDM and PixCell; and as a representation-alignment target that also tunes the VAE end-to-end for REPA-E. RAE and SVG remain the clean encoder-controlled tests of our claim, since only they share STREAM's latent space.

Evaluation. We evaluate with three natural-image-pretrained distribution metrics (FID [22], Kernel Inception Distance (KID) [6], CLIP Maximum Mean Discrepancy (CMMD) [28]) and three histopathology-aware variants FvD, KvD, vMMD, in which the underlying feature extractor is replaced by Virchow2 [78], another state-of-the-art histopathology VFM. LPIPS [76] is also used to measure reconstruction quality. STREAM, RAE and SVG are sampled at a matched number of function evaluations (NFE=50). All models are scored at 50k generated samples against the same real reference set.

5.2 Image Generation Results

As shown in Table 3, STREAM achieves the best gFID on both datasets (BRCA 6.16, COADREAD 7.68) against all other baselines. PixCell's design prevents unconditional generation: it always requires an input image (or a VFM embedding extracted from one) at inference. Forced to generate from its learned latent space alone, PixCell reports extremely high gFID values, consistent with the measured conditioning-dominance ratio ρ_{cond}=62%. ZoomLDM partially sidesteps this requirement by training a conditional diffusion model directly on the VFM embedding, allowing unconditional sampling without an external image; yet its gFID is higher than STREAM's. Both results indicate that the VAE-induced latent space is too semantically poor to support true generative modeling, even where reconstruction remains strong.

STREAM instead treats the VFM embedding as the latent space itself, without external conditioning, outperforming its histopathology-domain counterparts. However, STREAM's performance is not solely due to UNI being utilized as the latent space: pure RFM without STREAM's bridge × anisotropic-decoder design actually performs worse than ZoomLDM in both reconstruction and generation (Table 4, 45k decoder steps). STREAM also outperforms the VFM-latent baselines RAE and SVG, and because the UNI encoder is held fixed across the table, that margin belongs to the stochastic-bridge DiT and the anisotropic decoder rather than to the latent space. Furthermore, STREAM at half NFE (NFE=26) still beats RAE and SVG at the full NFE (NFE=50) (Appendix D.2).

Table 3: Generation results on BRCA and COADREAD $20\times$ patches (256×256), with all six methods on the same UNI encoder. All metrics lower-is-better; **best** bold, second-best underlined within each block. LPIPS is reconstruction-only.

BRCA							
Method	FID	KID	CMMD	LPIPS	FvD	KvD	vMMD
Reconstruction							
ZoomLDM	4.88	0.0040	0.2685	0.048	37.26	0.038	4.12
PixCell	2.88	0.0025	0.2590	**0.036**	18.11	0.018	2.04
STREAM	2.23	0.0019	**0.1509**	0.047	**9.91**	**0.008**	**0.98**
RAE	6.60	0.0058	1.0225	0.105	30.33	0.059	6.09
SVG	**2.16**	**0.0017**	0.8534	0.049	16.68	0.020	2.21
REPA-E	26.83	0.0211	1.4417	0.162	262.65	0.468	51.37
Generation							
ZoomLDM	7.43	0.0058	0.2611	–	196.41	0.194	21.01
PixCell	104.18	0.0984	0.6147	–	1298.00	1.410	169.38
STREAM	**6.16**	**0.0041**	**0.1282**	–	**78.04**	**0.048**	**5.16**
RAE	15.84	0.0126	1.0960	–	103.14	0.172	17.26
SVG	17.37	0.0149	0.8895	–	123.04	0.185	19.90
REPA-E	28.57	0.0224	1.4155	–	309.39	0.549	60.14

COADREAD							
Method	FID	KID	CMMD	LPIPS	FvD	KvD	vMMD
Reconstruction							
ZoomLDM	4.12	0.0033	0.1068	0.050	29.50	0.024	2.35
PixCell	2.58	0.0021	0.0796	**0.038**	**15.87**	**0.012**	**1.20**
STREAM	5.64	0.0043	0.0407	0.075	32.08	0.030	2.72
RAE	6.31	0.0050	0.0872	0.115	35.31	0.056	5.43
SVG	**1.81**	**0.0012**	**0.0304**	0.045	21.37	0.028	2.78
REPA-E	13.58	0.0096	0.2654	0.172	324.93	0.554	50.47
Generation							
ZoomLDM	8.09	0.0053	0.0613	–	139.63	0.097	9.52
PixCell	127.24	0.1132	0.9570	–	1742.85	2.145	218.70
STREAM	**7.68**	**0.0042**	**0.0293**	–	91.96	**0.055**	**5.10**
RAE	19.64	0.0144	0.1605	–	**91.74**	0.129	11.51
SVG	23.64	0.0189	0.1042	–	152.29	0.209	18.52
REPA-E	13.41	0.0090	0.2362	–	345.29	0.568	51.62

STREAM is not state of the art on reconstruction. SVG attains the best rFID on both datasets, and on COADREAD both PixCell and ZoomLDM also reconstruct better than STREAM does. However, that advantage does not carry over to generation, where these models have higher gFID than STREAM. STREAM therefore combines state-of-the-art generation with reasonable reconstruction, as generative models should strive to achieve high generation fidelity first, then aim for a minimal reconstruction trade-off. The histopathology-aware FvD, KvD and vMMD do not always rank methods the same way as the natural-image-pretrained FID, KID and CMMD, so we report both families. Regardless, STREAM ranks first on nearly all of the generative metrics. The same holds on the independent non-TCGA SPIDER-skin dataset, where all six methods are retrained from scratch on UNI (Appendix E.1).

We further evaluate STREAM-generated images for biological fidelity in two additional ways. First, we measure nuclear morphometry directly on the generated images, and STREAM reproduces the quantitative nuclear morphometry metrics of real tissue. Second, we test the prospective downstream application of STREAM, which is data augmentation for training self-supervised foundation models. A proof-of-concept encoder pretrained only on STREAM-generated samples of BRCA and COADREAD is on par with one pretrained on real tiles on clinically relevant downstream classification and gene expression prediction tasks (Appendix F).

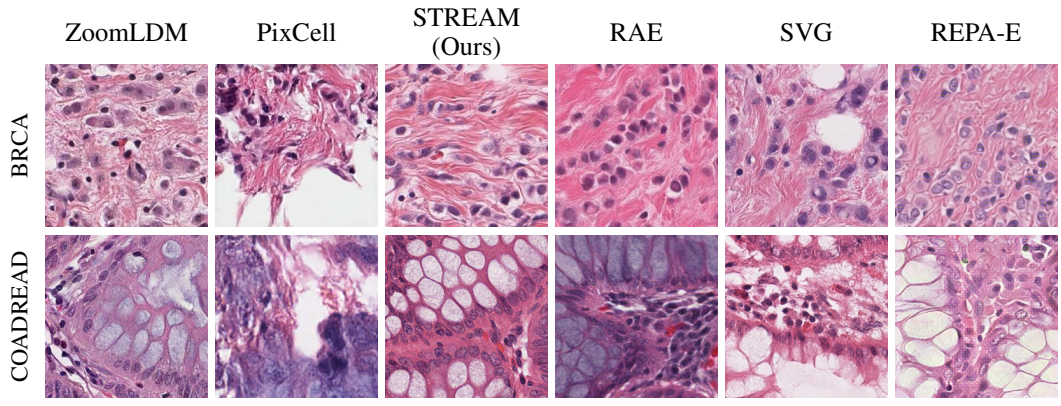


Figure 2: Generation samples from each trained model on BRCA and COADREAD.

For a qualitative check, Fig. 2 shows generation samples from ZoomLDM, PixCell, STREAM, RAE, SVG and REPA-E, selected by K-means clustering of ground-truth (GT) images in the Virchow2 feature space; for each cluster, we display the per-model sample whose Virchow2 feature is closest to the cluster centroid. Consistent with the gFID values, Fig. 2 shows that REPA-E and PixCell exhibit visible artifacts and structural features that do not represent real histopathology tissue, while the other baseline models produce globally plausible tissue but with visible color and

Table 4: Ablation on BRCA (UNI encoder) at 45k decoder steps.

Bridge	Decoder	rFID ↓	gFID ↓
✗	Isotropic	6.51	9.07
✗	Anisotropic (Ours)	5.02	8.27
✓ (Ours)	Isotropic	6.51	9.37
✓ (Ours)	Anisotropic (Ours)	**3.52**	**6.86**

Table 5: Anisotropic decoder directional preference on BRCA, n=2,000 tiles.

R_{dec}	Isotropic	Anisotropic (Ours)
STREAM (k=64)	0.42	3.41
STREAM (k^*=236)	1.12	2.31
No bridge (k=64)	0.62	2.50
No bridge (k^*=232)	1.28	2.01

texture artifacts when representing the stromal region of the BRCA image and the glandular region of the COADREAD image. In contrast, STREAM produces realistic-looking histopathology images with well-defined nuclei morphology and accurate textural, structural, and color detail. Additional reconstruction and generation samples for all six models are provided in Appendix I.

6 Ablation Studies

Ablation on STREAM design. All ablation studies are performed on the BRCA dataset with identical training hyperparameters, except that the anisotropic decoder is trained for 45k steps. Table 4 isolates the two core STREAM contributions. Anisotropic noise alone (without the bridge) yields a meaningful improvement: rFID drops from 6.51 to 5.02 and gFID from 9.07 to 8.27, consistent with allocating noise budget away from the high-response directions U_H. The bridge alone with isotropic noise marginally degrades gFID (9.07 → 9.37) while leaving rFID unchanged. Yet the full combination produces a superadditive effect: rFID drops to 3.52 and gFID to 6.86, far exceeding the sum of the individual contributions. STREAM's superadditive design ties the decoder's noise to the left-singular basis of the velocity-field Jacobian, which reverses RAE's isotropic-noise rFID–gFID asymmetry [77] and delivers the joint improvement of Table 4. Diffusion in high-dimensional latents is commonly held to trade generation for reconstruction [71], yet STREAM generates directly in UNI's 1024-dimensional VFM latent space and still improves both metrics at once.

The superadditive effect has an empirical explanation in terms of directional robustness. Given (CT) of Proposition 3, the bridge's transport perturbation is asymptotically $O(\sigma_{\max}^2)$, and by Theorem 1(iii) the resulting marginal velocity field is smooth at intermediate t, yet the bridge changes the spectral structure that the decoder's cache is subsequently extracted from; under an isotropic decoder, which has no directional structure to exploit, that change does not help and gFID drifts from 9.07 to 9.37. The anisotropic decoder turns that change into a gain by spending its robustness budget on the low-response subspace U_L ($\sigma_L = 0.02$) while holding $\sigma_H = 0.002$ on U_H. We do not claim that the generator's latent error concentrates in U_L (Appendix G.4). In both subspaces the decoder trains under more noise than the generation error it later has to absorb. The directional-preference probe below shows which subspace the decoder ends up most sensitive to.

Anisotropic decoder directional preference. At the decoder level, we directly test the spectral mechanism via a per-token directional Lipschitz contrast on 2,000 BRCA tiles, run under two partitions, k=64 and each model's own τ=0.90 energy cutoff. For each image we sample one unit-norm tangent direction in U_H and one in the orthogonal complement U_L (within the cached per-(centroid, token) basis), perturb z at matched per-token tangent norm $\epsilon \in \{0.005, 0.01, 0.016, 0.02\}$, decode, and measure LPIPS sensitivity. The probe range brackets the decoder's training noise envelope. The anisotropic decoder trains with $\sigma_L = 0.02$ as the maximum low-energy magnitude, so the largest probe matches the training scale and the smaller values stay strictly inside it. Per-image polyfit slopes give per-direction Lipschitz constants $L^{\text{dec}}_{U_H}, L^{\text{dec}}_{U_L}$; their ratio $R_{\text{dec}} = L^{\text{dec}}_{U_H}/L^{\text{dec}}_{U_L}$ characterizes the decoder's directional preference, with values > 1 indicating U_H-preference and < 1 indicating U_L-preference. With or without the bridge, the isotropic decoder favors U_L more than the anisotropic decoder does (Table 5), most likely because of data structure. Along-manifold tangent variations produce visually meaningful image changes that the decoder must preserve, while reconstruction largely smooths out manifold-normal variations. Anisotropic training tilts the decoder's preference toward U_H in every spherical configuration we probe. Regardless of where the isotropic decoder starts, anisotropic training raises R_{dec}: from 0.42 to 3.41 at k=64 (8.1×) and from 1.12 to 2.31 at the deployed k^*=236 (2.1×). The rise is larger with the bridge than without it at both splits, matching the superadditive gains of Table 4. The same probe run on the Euclidean counterpart of

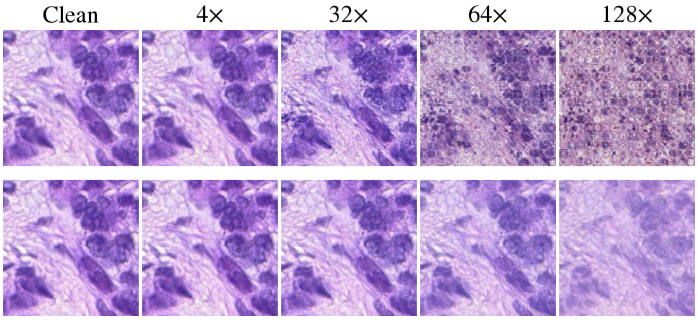


Figure 3: Pixel response (BRCA) of the anisotropic (upper row) and isotropic (lower row) decoders to a top-64 subspace perturbation for the clean latent and four amplitudes.

Table 6: Encoder-dependence ablation on BRCA: domain-mismatched DINOv2-L.

Decoder	rFID ↓	gFID ↓
Isotropic	7.67	11.05
Anisotropic (Ours)	7.71	11.22

RAE barely moves the ratio, so without the spherical geometry there is no U_H/U_L split for the decoder to exploit (Appendix G.5). The deployed $\sigma_L = 0.02$ sits at the gFID optimum of a decoder noise ablation on BRCA (Appendix G.1).

Ablation on the training loss. To our knowledge, no prior RFM method regresses endpoints in the ambient chord metric. At the full 90k decoder schedule, STREAM's chord loss with x-prediction beats standard RFM (geodesic loss, v-prediction) on both rFID and gFID. Aligned with the current literature [36], x-prediction with the chord loss performs the best, while both mismatched pairings of loss and target give the worst gFID (Appendix E.4). The chord loss performs better because it shares the per-sample geodesic minimizer but avoids the $\arccos$ singularity (Appendix C.2).

Ablation on encoder dependence. To test whether STREAM's advantages hold on encoders other than UNI, we re-run the full pipeline with Virchow2 [78], another state-of-the-art histopathology VFM. With Virchow2 the anisotropic decoder gains 4.91 rFID and 4.27 gFID over its isotropic control (Appendix H.3). The design therefore survives a change of histopathology VFM and is not tuned to UNI's particular spectrum. However, replacing UNI with DINOv2-L, a domain-mismatched natural-image VFM, removes the isotropic-to-anisotropic gain of UNI and Virchow2 (Table 6).

The DiT does not account for this: it matches DINOv2-L's latent distribution about as closely as the UNI DiT matches UNI's (latent FD 0.032 vs. 0.030, Appendix H.1). The missing gain shows already in reconstruction, which involves no generated latent: anisotropic noise leaves DINOv2-L's rFID unchanged ($7.67 \rightarrow 7.71$) but lowers UNI's by 2.99 (Table 4). We therefore place the failure on the decoder side, where the spectral noise shaping does not help on DINOv2-L's natural-image features. An in-domain histopathology VFM remains the reliable choice (Appendix H.4). We do not yet have a statistic that predicts in advance whether a new encoder benefits; the isotropic-versus-anisotropic decoder contrast is the empirical test. Under a perturbation inside the top-64 subspace (U_H at k=64), the anisotropic decoder separates from the isotropic one and responds more strongly (Fig. 3), a pixel-space view of the decoder sensitivity of Table 5 behind the superadditive effect. As the perturbation amplitude grows from $4\times$ to $128\times$, the anisotropic decoder's output changes visibly from $32\times$ onward while the isotropic decoder's changes far less; the anisotropic decoder tracks latent movement along U_H instead of smoothing it out. Both decoders see the same perturbed latent and differ only in training noise, so the difference reflects the U_H-preference of Table 5.

7 Conclusion

We present STREAM, an RFM framework for unconditional histopathology generation in VFM latent spaces, pairing a stochastic bridge on $\mathcal{S}^{d-1}$ with a spectral-informed anisotropic decoder. The two components combine superadditively, and under a single shared UNI encoder STREAM attains state-of-the-art generation quality on three different histopathology datasets. Main limitations that remain are that Theorem 1 holds at the population level and that our account of the superadditive mechanism is mostly empirical (Appendix G.4). Future work includes transferring STREAM to other imaging modalities, extending the method to conditional generation, and evaluating STREAM-generated images far more extensively, at the level of an actual clinical setting.

Acknowledgments and Disclosure of Funding

This work was supported by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, South Korea). All authors are employees of DEEPNOID Inc. and declare no other competing interests.

References

[1] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

[2] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*. Springer, 1985.

[3] Elias Baumann, Bastian Dislich, Josef Lorenz Rumberger, Iris D. Nagtegaal, Maria Rodriguez Martinez, and Inti Zlobec. HoVer-NeXt: A fast nuclei segmentation and classification pipeline for next generation histopathology. In *Medical Imaging with Deep Learning (MIDL)*, volume 250 of *Proceedings of Machine Learning Research*, pages 61–86, 2024.

[4] Jean-David Benamou and Yves Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[5] Mahesh Bhosale, Abdul Wasi, Yuanhao Zhai, Yunjie Tian, Samuel Border, Nan Xi, Pinaki Sarder, Junsong Yuan, David Doermann, and Xuan Gong. PathDiff: Histopathology image synthesis with unpaired text and mask conditions. In *ICCV*, 2025. arXiv:2506.23440.

[6] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

[7] Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[8] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301–1309, 2019.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[10] Hun Chang, Byunghee Cha, and Jong Chul Ye. HAE: Hyperspherical autoencoder for high-fidelity image reconstruction and generation. *arXiv preprint arXiv:2601.22904*, 2026.

[11] Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In *ICLR*, 2026.

[12] Hao Chen, Yujin Han, Fangyi Chen, et al. Masked autoencoders are effective tokenizers for diffusion models. In *ICML*, 2025.

[13] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[14] Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *ICLR*, 2024.

[15] Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[17] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorber, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[19] Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

[20] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *CVPR*, 2024.

[21] Johannes Hertrich, Antonin Chambolle, and Julie Delon. On the relation between rectified flows and optimal transport. *arXiv preprint arXiv:2505.19712v3*, 2025.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[24] Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models. *arXiv preprint arXiv:2411.02817*, 2024.

[25] Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. *arXiv preprint arXiv:2506.06231*, 2025.

[26] Anne F. J. Janssen, Sophia Y. Breusegem, and Delphine Larrieu. Current methods and pipelines for image-based quantitation of nuclear shape and nuclear envelope abnormalities. *Cells*, 11 (3):347, 2022.

[27] Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. In *NeurIPS*, 2024.

[28] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] kaiko.ai, Nanne Aben, Edwin D. de Jong, Ioannis Gatopoulos, Nicolas Känzig, Mikhail Karasikov, Axel Lagré, Roman Moser, Joost van Doorn, and Fei Tang. Towards large-scale training of pathology foundation models. *arXiv preprint arXiv:2404.15217*, 2024.

[30] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *ICML*, 2025.

[31] Amandeep Kumar and Vishal M. Patel. Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. In *ECCV*, 2026.

[32] Minh-Quan Le, Alexandros Graikos, Srikar Yellapragada, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras. ∞-Brush: Controllable large image synthesis with diffusion models in infinite dimensions. In *ECCV*, 2024.

[33] John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2nd edition, 2018.

[34] Junho Lee, Kwanseok Kim, and Joonseok Lee. Geometry-aware image flow matching. In *ICML*, 2026.

[35] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *ICCV*, 2025.

[36] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

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

[38] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[39] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

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

[41] Grégoire Loeper. On the regularity of solutions of optimal transportation problems. *Acta Mathematica*, 202(2):241–283, 2009.

[42] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[43] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024.

[44] Xi-Nan Ma, Neil S. Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Archive for Rational Mechanics and Analysis*, 177(2):151–183, 2005.

[45] Seonghui Min, Hyun-Jic Oh, and Won-Ki Jeong. Co-synthesis of histopathology nuclei image-label pairs using a context-conditioned joint diffusion model. In *ECCV*, 2024. doi: 10.1007/978-3-031-72624-8_9.

[46] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. SPIDER: A comprehensive multi-organ supervised pathology dataset and baseline models. *arXiv preprint arXiv:2503.02876*, 2025.

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

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

[49] Bruno Pelletier. Nonparametric regression estimation on closed Riemannian manifolds. *Journal of Nonparametric Statistics*, 18(1):57–67, 2006.

[50] Kai Qiu, Xiang Li, Hao Chen, Jason Kuen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Image tokenizer needs post-training. *arXiv preprint arXiv:2509.12474*, 2025.

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

[52] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

[53] Qingyu Shi, Size Wu, Jinbin Bai, Kaidong Yu, Yujing Wang, Yunhai Tong, Xiangtai Li, and Xuelong Li. RecTok: Reconstruction distillation along rectified flow. In *CVPR*, 2026.

[54] Oriane Simeoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[55] Yang Song, Jasper Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[56] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. UniLiP: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.

[57] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012. doi: 10.1038/nature11252.

[58] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[59] Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27:775–784, 2021.

[60] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014.

[61] Max-K. von Renesse and Karl-Theodor Sturm. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on Pure and Applied Mathematics*, 58(7):923–940, 2005.

[62] Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31 (2):609–617, 2025.

[63] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

[64] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025.

[65] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, Jason Wei, and Saeed Hassanpour. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine (AIME)*, volume 12721 of *Lecture Notes in Computer Science*, pages 11–24, 2021.

[66] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. A vision-language foundation model for precision oncology. *Nature*, 638(8051): 769–778, 2025.

[67] Conghao Xiong, Zhengrui Guo, Zhe Xu, Yifei Zhang, Raymond Kai-yu Tong, Si Yong Yeo, Hao Chen, Joseph J. Y. Sung, and Irwin King. Exploiting low-dimensional manifold of features for few-shot whole slide image classification. In *ICLR*, 2026.

[68] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[69] Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, and Chao Chen. TopoCellGen: Generating histopathology cell topology with a diffusion model. In *CVPR*, 2025.

[70] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. *arXiv preprint arXiv:2507.15856*, 2025.

[71] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

[72] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. PathLDM: Text conditioned latent diffusion model for histopathology. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024.

[73] Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Tarak Nath Nandi, Karen Bai, Beatrice S. Knudsen, Tahsin Kurc, Rajarsi R. Gupta, Prateek Prasanna, Ravi K. Madduri, Joel Saltz, and Dimitris Samaras. PixCell: A generative foundation model for digital histopathology images. *arXiv preprint arXiv:2506.05127*, 2025.

[74] Srikar Yellapragada, Alexandros Graikos, Kostas Triaridis, Prateek Prasanna, Rajarsi R. Gupta, Joel Saltz, and Dimitris Samaras. ZoomLDM: Latent diffusion model for multi-scale image generation. In *CVPR*, 2025.

[75] 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 *ICLR*, 2025.

[76] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[77] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

[78] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

Mathematical Notation

Symbols carried through the main text. A symbol marked † is overloaded within the main text; both senses are given in the same row. Symbols confined to a single appendix are defined where they appear and are not repeated here.

Symbol	Meaning
$\mathcal{S}^{d-1}$, $(\mathcal{S}^{d-1})^N$	Unit hypersphere $\{x \in \mathbb{R}^d : \|x\| = 1\}$, the VFM latent manifold, and the product of N per-token spheres carrying $g = \bigoplus_n g_n$ (Sections 4 and 4.1)
d, N, $n^\dagger$	Embedding dimension per token (d=1024 for UNI), token count per image (N=256), so the total tangent dimension is $N(d-1) = 261{,}888$. n is the token index and, in Eq. (7), the ambient noise vector added to a token (Section 4.1)
Exp_x, Log_x, d_g, Ω	Riemannian exponential map, logarithmic map and geodesic distance on $\mathcal{S}^{d-1}$, and the endpoint angle $\Omega = d_g(x_0, x_1) \le \pi$ of the geodesic conditional path (Appendix B.1 and Eq. (2))
$t^\dagger$	Flow time, t=0 at the source and t=1 at the data; also the fixed anchor t=0.99 at which the spectral basis is cached (Definition 1 and Section 4.3)
p_0, p_1, p_t	Source (uniform), data and interpolating distributions on $(\mathcal{S}^{d-1})^N$; $p_{\text{data}} = E_{\#}(p_{\text{real}})$ and p_θ is the generated feature distribution (Section 4)
$\pi^\dagger$	Source–data coupling ($p_0 \otimes p_1$ in Definition 1, a general coupling in Proposition 1); also the circle constant, in $\sin(\pi t)$ and the sphere diameter (Definition 1)
x_0, x_1, μ_t, $x_t{}^\dagger$, X_t	Source and data endpoints on one sphere, the SLERP interpolant, the bridge-perturbed interpolant, and that interpolant as a random variable; in Eq. (3) x_t is instead the unperturbed SLERP point (Eq. (4))
$\epsilon^\dagger$, ϵ_t, Π_μ	Tangent Gaussian noise of the bridge, $\epsilon \sim \mathcal{N}(0, \Pi_{\mu_t})$ with projector $\Pi_\mu = I - \mu\mu^\top$; ϵ is also the tangent probe magnitude of the directional-Lipschitz test, and is distinct from the time clamp $\epsilon_t = 10^{-5}$ (Eq. (4))
$\sigma(t)$, $\sigma_{\max}$	Bridge noise schedule $\sigma_{\max}\sin(\pi t)$ and its peak amplitude, deployed $\sigma_{\max} = 0.01$ (Definition 1)
u_t, v_t, v_θ	Conditional velocity along the known path, the marginal velocity field it induces, and the learned velocity field (Section 4)
E, G_θ, D_ϕ, $\mathcal{X}$	Frozen VFM encoder $\mathcal{X} \to (\mathcal{S}^{d-1})^N$, DiT generator, ViT decoder $(\mathcal{S}^{d-1})^N \to \mathcal{X}$, and the image space $\mathcal{X} = \mathbb{R}^{H\times W\times 3}$ (Sections 4.1 and 4.4)
$\theta^\dagger$, ϕ	Generator and decoder parameters; θ is also the geodesic angle argument of Log_x (Eqs. (3) and (12))
f_θ, $\hat{x}_1$	The DiT read as an endpoint (x-prediction) network, and its normalized output $f_\theta(x_t,t)/\|f_\theta(x_t,t)\|$ (Eq. (5))
$\mathbf{Z}_0$, $\mathbf{Z}_1$, $\hat{\mathbf{Z}}_1$	All-token feature tensors on $(\mathcal{S}^{d-1})^N$: source, clean data, and generated (Section 4.1)
$z^\dagger$, $\tilde{z}$, z_{rt}	A single clean VFM token on $\mathcal{S}^{d-1}$, the noised and renormalized token $(z+n)/\|z+n\|$, and its round-trip re-encoding $E(D_\phi(\tilde{z}))$. Latents are z and generic points of $\mathcal{S}^{d-1}$ are x; z also names the conditioning triple (x_0, x_1, ϵ) of the generic CFM loss (Eq. (7))
$I^\dagger$	Input training image, reconstructed as $\hat{I} = D_\phi(\tilde{z})$; also the identity matrix, and the mutual information $I(X;C)$ (Section 4.1)
$\mathcal{L}_{\text{RFM}}$, $\mathcal{L}_{\text{chord}}$, $\mathcal{L}_{\text{dec}}$	RFM loss, the deployed chord loss on normalized endpoints, and the four-component decoder loss with weights λ_{ℓ_1}, λ_{LPIPS}, $\lambda_{\cos}$, $\omega_G\lambda_{\text{adapt}}$ (Eqs. (3), (5) and (7))
J	Per-token tangent-projected Jacobian of the velocity field, $\nabla_z v_\theta(z,t)$ (Section 4.3)
U_H, U_L	High-response (top-k^*) and low-response left singular subspaces of J; H and L label the two throughout. Unrelated to the $U[0,1]$ of Eq. (3) (Section 4.3)
s_i, τ, k, k^*	Singular values of J; the spectral energy threshold, deployed 0.90; and the retained-dimension count $k^* = \min\{k : \sum_{i\le k} s_i^2 / \sum_i s_i^2 \ge \tau\}$. k is a generic retained-subspace dimension (k=64 in the alignment and Lipschitz probes); k^* is the count that τ fixes, not the cutoff itself; it is taken on the median spectrum of the cache and is model-specific. Singular values are written s_i, never σ_i (Section 4.3)
$\Sigma_{\text{noise}}(z)$, σ_H, σ_L	Anisotropic decoder noise covariance $\sigma_H^2 U_H U_H^\top + \sigma_L^2 U_L U_L^\top$ and its two magnitudes, deployed 0.002 and 0.02 (Eq. (6))
K, m, $\bar{z}_c$	Number of k-means centroids of the spectral cache (K=50), random tangent probes per (centroid, token) (m=600), and the c-th centroid. K separately denotes the Vendi similarity kernel (Section 4.3)
R_{eff}, R_{dec}, A	Effective rank of the feature covariance, which is not the Jacobian cut k^* (Section 3.2); the decoder's directional-Lipschitz ratio $L^{\text{dec}}_{U_H}/L^{\text{dec}}_{U_L}$ (Section 6); and the principal-angle subspace alignment $\|Q_1^\top Q_2\|_F^2/k \in [0,1]$ (Appendix H.2)
ρ_{cond}, $H(X)$, C	Conditioning dominance ratio $I(X;C)/H(X)$, the marginal output entropy, and the conditioning variable $C = E_{\text{VFM}}(I)$ (Eq. (1))
f_{angular}, Acc_{dir}, Acc_{mag}	Angular fraction of the discriminative between-class distance, and 1-NN accuracy on the normalized direction alone and on the norm $\|z\|$ alone (Appendix A.4)

A Conditioning-Dominated Diversity Analysis

We provide the full conditioning-dominated-diversity analysis summarized in Section 3.1. Throughout, "conditioning-dominated diversity" names the regime in which $\rho_{\text{cond}} = I(X;C)/H(X)$ is close to 1, i.e. almost all of a model's output diversity is attributable to variation in the conditioning signal C rather than to its latent space. It is a statement about where the diversity comes from, not a claim that the models produce degenerate images.

A.1 Entropy Estimation via Vendi Score

For VFM-conditioned models, the output diversity decomposes as $H(X) = H(X \mid C) + I(X;C)$, where $C = E_{\text{VFM}}(I)$ is the conditioning embedding. We estimate these quantities using the Vendi Score of order 1 [19], which computes $\text{Vendi}_1(K) = \exp(H_1(\lambda(K)))$ where $\lambda(K)$ are the normalized eigenvalues of the similarity kernel matrix K and H_1 is Shannon entropy. For a set of samples $\{X_j\}$, $\log \text{Vendi}_1$ estimates the Shannon entropy of the kernel-induced output distribution.

Key property. Because both baselines are sampled with deterministic samplers that inject no noise during the reverse process, the only source of randomness is the initial latent noise $z_T \sim \mathcal{N}(0, I)$. The entropy we estimate is therefore model-induced diversity (the pushforward entropy of $\mathcal{N}(0, I)$ through $f_\theta(\cdot, c)$), not sampler noise. This avoids step-wise noise inflation and confounding stochasticity from the sampling procedure.

Important caveats. The quantities we estimate are not: (i) the true data entropy $H(p_{\text{data}}(X \mid C))$; (ii) the training objective entropy; or (iii) Shannon mutual information $I(X;C)$ in pixel space. Rather, we estimate the conditional entropy of the generative model's output distribution, under a fixed inference procedure, measured in VFM (UNI) embedding space. This is a valid and informative diagnostic because UNI embeddings are approximately information-preserving for tissue-level semantics, which is the relevant notion of diversity in histopathology. The difference $I(X;C) = H(X) - H(X \mid C)$ is a kernel-induced Shannon mutual information, which is monotonic in diversity and comparable across models evaluated with the same kernel and embedding space.

A.2 Measurement Protocol and Discussion

Step 1: Clustering. We apply Euclidean k-means (FAISS) on BRCA UNI embeddings ($\sim$500k samples) to obtain conditioning clusters. We evaluate $K \in \{5, 8, 10, 12, 15, 20, 30\}$, selecting $K^* = 10$ between the elbow and silhouette candidates by the higher Adjusted Rand Index stability across five k-means seeds. Cluster weights $w_k = |C_k|/N_{\text{total}}$ reflect the empirical conditioning distribution.

Step 2: Conditional entropy $H(X \mid C)$. For each cluster k, we sample conditioning vectors $c_{k,m}$ and generate multiple images per condition by varying z_T. The per-condition Vendi Score estimates $H(X \mid C = c_{k,m})$. Averaging over conditions with cluster weights gives $\hat{H}(X \mid C) = \sum_k w_k \cdot \mathbb{E}_m[H(X \mid C = c_{k,m})]$.

Step 3: Marginal entropy $H(X)$. We pool all generated samples across all conditions and compute a single Vendi Score on the full set. This approximates $H(X)$ via Monte Carlo marginalization over C.

Step 4: Mutual information. $I(X;C) = H(X) - H(X \mid C)$, normalized as $\rho_{\text{cond}} = I(X;C)/H(X)$.

Step 5: Condition–output spectral gap. We compute SPEC-diff [25], the operator norm of $(K_C - K_X)/M$, with K_C and K_X the cosine-kernel matrices of M conditioning embeddings and of the UNI embeddings of the M images generated from them. Low CVS combined with low SPEC-diff is a signature of conditioning-dominated diversity.

Discussion. The decomposition $H(X) = H(X \mid C) + I(X;C)$ separates domination from desirable conditioning dependence directly. Desirable dependence means the model uses C and still produces varied outputs once C is held fixed, so $H(X \mid C)$ stays large and ρ_{cond} small. Domination means nearly all the diversity is attributable to C varying, so $H(X \mid C)$ falls away and ρ_{cond} approaches 1. Table 1 measures the latter on BRCA in UNI space: ρ_{cond} is 75% for ZoomLDM and 62% for PixCell, with $H(X \mid C)$ of 1.28 and 1.81 against totals of 5.17 and 4.78. Equivalently,

latent utilization is $1 - \rho_{\text{cond}} = H(X \mid C)/H(X)$, i.e. 25% and 38%, and output diversity at fixed condition is $H(X \mid C)$; both quantities are already in Table 1 under other names.

Removing C separates the two regimes operationally. PixCell cannot sample without being handed a VFM embedding, and ZoomLDM only sidesteps that by learning a second generative model over the embedding. Forced to sample from its own latent space, PixCell reaches 104.18 gFID on BRCA against 6.16 for STREAM (Table 3), so its latent space alone does not support generation.

STREAM's marginal-entropy check. STREAM is unconditional, so ρ_{cond}, $I(X;C)$ and $H(X \mid C)$ do not apply to it; $H(X)$ needs only samples, under the marginal-entropy estimator of Step 3 above. STREAM's $H(X) = 5.26$ from noise alone exceeds either baseline's total of 5.17 and 4.78: it generates more variety with no conditioning signal than they achieve with their condition free to vary.

A.3 Linear Probing: SPIDER-breast Dataset

The linear probing evaluation uses SPIDER-breast [46] with an 11-class binary split into Benign/Proliferative (6 classes, 26,664 patches) and Malignant (5 classes, 27,164 patches); class-level breakdown in Table 7. This probe measures the encoder's representation quality and is distinct from the downstream evaluation of Appendix F, which probes encoders pretrained from scratch on STREAM-generated images. A linear classifier is trained on frozen embeddings from each model/encoder; AUROC with a 95% bootstrap confidence interval (CI) is reported in Table 8.

Table 7: SPIDER-breast class composition. DCIS is ductal carcinoma in situ and NST is no special type.

Class	Patches
Benign/Proliferative (6 classes)	
Adenosis	2,899
Sclerosing adenosis	3,423
Fibrocystic changes	5,027
Typical ductal hyperplasia	5,546
Fibroadenoma	5,243
Benign phyllodes	4,526
Subtotal	*26,664*
Malignant (5 classes)	
DCIS low-grade	5,017
DCIS high-grade	5,632
Invasive NST carcinoma	6,142
Lobular invasive carcinoma	5,102
Malignant phyllodes	5,271
Subtotal	*27,164*
Total	**53,828**

Table 8: Linear-probe AUROC (95% bootstrap CI) on SPIDER-breast.

Encoder	d	AUROC [95% CI]
UNI2-h	1536	**0.997** [0.996, 0.998]
UNI	1024	0.995 [0.993, 0.996]
DINOv3-L	1024	0.937 [0.931, 0.942]
PixCell VAE	16	0.640 [0.628, 0.653]
ZoomLDM VAE	3	0.624 [0.611, 0.637]

A.4 1-NN Direction vs. Magnitude Perturbation Protocol

This subsection details the protocol behind the Acc_{dir}, Acc_{mag}, and f_{angular} columns of Table 2, as referenced from the angular-dominance argument in Section 3.1. The setup is shared with Appendix A.3: all five models are evaluated on the SPIDER-breast binary task (benign/proliferative vs. malignant; $N_{\text{test}} = 6{,}839$) using frozen embeddings, with no fine-tuning.

1-NN leave-one-out (LOO) baseline. For each test embedding z_i with label y_i, we find its nearest same-class and nearest different-class neighbor in the test set under Euclidean distance (excluding z_i itself), and count a correct hit iff $d(z_i, \text{NN}_{\text{same}}) < d(z_i, \text{NN}_{\text{diff}})$. We report three accuracies on the same test split:

- Acc_{full}: 1-NN on the raw embedding z.
- Acc_{dir} (Table 2): 1-NN on the ℓ_2-normalized direction $\hat{z} = z/\|z\|$, isolating the angular component.

- Acc_{mag} (Table 2): 1-NN on the 1-D scalar $\|z\|$, isolating the radial component.

The gap $\text{Acc}_{\text{full}} - \text{Acc}_{\text{dir}}$ is a direct estimate of how much classification value the magnitude adds beyond direction; for all three VFMs this gap is below 0.005, i.e. direction alone reproduces the full 1-NN accuracy.

Variance decomposition of between-class distance. For each pair (z_i, z_j), the squared Euclidean distance factors as

$$\|z_i - z_j\|^2 = \underbrace{(\|z_i\| - \|z_j\|)^2}_{\Delta_{\text{radial}}(i,j)} + \underbrace{2\,\|z_i\|\,\|z_j\|\,(1 - \cos\theta_{ij})}_{\Delta_{\text{angular}}(i,j)}, \tag{8}$$

where θ_{ij} is the angle between z_i and z_j. Averaging each component over the between-class (B) and within-class (W) pairs of the test set yields the angular fraction of discriminative distance

$$f_{\text{angular}} = \frac{\overline{\Delta}^{\text{B}}_{\text{angular}} - \overline{\Delta}^{\text{W}}_{\text{angular}}}{\left(\overline{\Delta}^{\text{B}}_{\text{angular}} - \overline{\Delta}^{\text{W}}_{\text{angular}}\right) + \left(\overline{\Delta}^{\text{B}}_{\text{radial}} - \overline{\Delta}^{\text{W}}_{\text{radial}}\right)}, \tag{9}$$

which we report in Table 2. All three VFMs satisfy $f_{\text{angular}} > 0.97$; the only model well below 1.0 is ZoomLDM VAE at 0.813. This is consistent with its 32.4% coefficient of variation of $\|z\|$ and its d=3 representation, in which the radial direction carries a meaningful share of the total information budget.

Angular perturbation stress test. To probe whether the 1-NN signal genuinely lives in the angular structure, we rotate each $\hat{z}_i$ in a random unit tangent direction by an angle θ and recompute 1-NN LOO after restoring each original norm $\|z_i\|$, so that $\cos\theta = 1.00$ returns Acc_{full}. We use two methods, additive Gaussian followed by re-normalization and exponential-map tangent rotation, and report the tangent method (the two agree within a small margin except for UNI below $\cos\theta = 0.50$, which the additive method cannot reach). We sweep $\cos\theta \in \{1.00, 0.80, 0.50, 0.30, 0.10, 0.00\}$ over three random seeds; Table 9 reports the means. UNI retains 96% accuracy up to $\cos\theta = 0.50$ (a 60° rotation), UNI2-h follows the same trend with a slightly steeper drop, DINOv3-L degrades smoothly from 0.929 down to chance, PixCell VAE collapses to chance by $\cos\theta = 0.80$, and ZoomLDM VAE falls to 0.592 there and plateaus at 0.583, above chance throughout. The VFM clusters thus enjoy angular separation on the order of tens of degrees. This is compatible with the bridge perturbation in Section 4.2, which operates well inside that margin.

Radial perturbation stress test. We replace $\|z_i\|$ by $\|z_i\| \cdot (1 + \tau \cdot \varepsilon_i)$ with $\varepsilon_i \sim \mathcal{N}(0,1)$ i.i.d. and re-evaluate 1-NN LOO; we sweep $\tau \in \{0.0, 0.2, 1.0, 2.0, 5.0, 10.0\}$ over three seeds (Table 10). UNI and UNI2-h hold accuracy ≥ 0.98 even at $\tau = 10$ (a $\pm 1000\%$ norm perturbation). A control check normalizes the perturbed embeddings before computing 1-NN; under that normalization the radial-perturbation curves are perfectly flat at Acc_{dir} for every model and every τ, confirming that the radial noise is entirely undone by re-projection onto $\mathcal{S}^{d-1}$.

Table 9: Angular perturbation stress test on SPIDER-breast (1-NN LOO accuracy under tangent rotation, mean over 3 seeds; $\cos\theta = 1.00$ recovers the unperturbed Acc_{full}). Chance level is 0.50.

Model	$\cos\theta$=1.00	0.80	0.50	0.30	0.10	0.00
UNI	0.993	0.992	0.960	0.674	0.521	0.506
UNI2-h	0.994	0.993	0.894	0.624	0.502	0.492
DINOv3-L	0.929	0.720	0.549	0.491	0.490	0.490
PixCell VAE	0.834	0.507	0.496	0.497	0.498	0.500
ZoomLDM VAE	0.726	0.592	0.585	0.587	0.586	0.583

Interpretation. The joint pattern is unambiguous: ($\text{Acc}_{\text{dir}} \approx \text{Acc}_{\text{full}}$) and ($f_{\text{angular}} > 0.97$) and (radial-perturbation curve flat at ≥ 0.98 for UNI and UNI2-h) and (normalized-radial control flat at Acc_{dir}) all together imply that the class-discriminative information in all three VFMs is carried almost entirely by the angular structure on $\mathcal{S}^{d-1}$. ZoomLDM VAE is the sole counter-example: its $f_{\text{angular}} = 0.81$ and the fact that ℓ_2-normalization drops its 1-NN accuracy by 11.7 pp ($0.726 \to 0.609$) show that magnitude carries non-redundant class information when the latent budget is as small as d=3. For the histopathology VFMs we use throughout STREAM, however, every diagnostic points to the sphere as the correct ambient manifold for generative modeling. That is exactly the structural premise the RFM formulation in Section 4.2 requires.

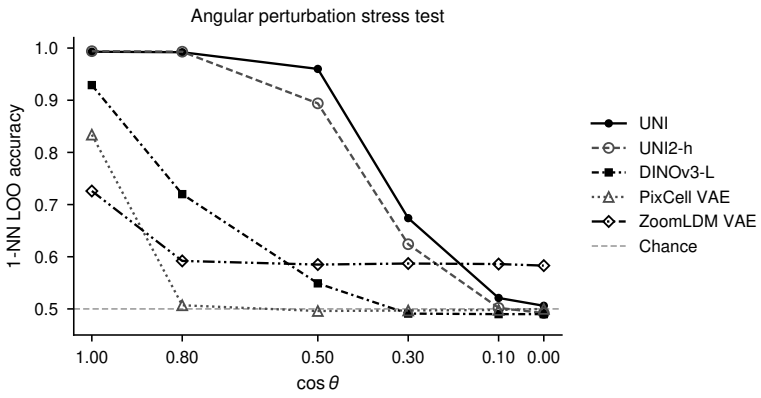


Figure 4: Angular perturbation stress test on SPIDER-breast: the means of Table 9 plotted against $\cos\theta$. Chance level is 0.50.

Table 10: Radial perturbation stress test on SPIDER-breast (1-NN LOO accuracy under $\|z\| \mapsto \|z\|(1+\tau\varepsilon)$, $\varepsilon \sim \mathcal{N}(0,1)$, mean over 3 seeds; $\tau = 0$ recovers Acc_{full}). UNI and UNI2-h hold ≥ 0.98 at $\tau = 10$.

Model	τ=0.0	0.2	1.0	2.0	5.0	10.0
UNI	0.993	0.990	0.985	0.985	0.983	0.985
UNI2-h	0.994	0.992	0.989	0.988	0.989	0.988
DINOv3-L	0.929	0.899	0.836	0.822	0.818	0.820
PixCell VAE	0.834	0.759	0.698	0.686	0.682	0.683
ZoomLDM VAE	0.726	0.617	0.606	0.600	0.597	0.598

A.5 Intrinsic Geometry: Extended Analysis

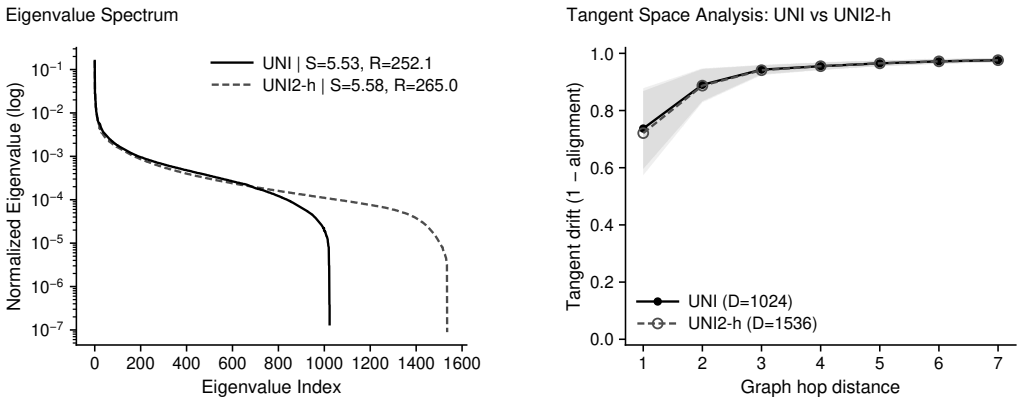
We extend the geometric characterization of Xiong et al. [67] to BRCA at scale ($\sim$1M 20$\times$ patches) for UNI (d=1024) and UNI2-h (d=1536), summarized in Fig. 5. Spectral analysis of the ℓ_2-normalized feature covariance reveals a high covariance effective rank, $R_{\text{eff}} = 252$ and 265 respectively, with von Neumann entropy $S_1 = 5.53$ and 5.58. This is consistent with the result of Xiong et al. [67], $R_{\text{eff}} = 235.5$ for UNI on Camelyon16, confirming that histopathology-specific self-supervised pretraining distributes semantic information across hundreds of dimensions. Tangent space analysis (k=12, intrinsic dimension d_s=10) shows drift rising steeply from 0.737 ± 0.141/0.721 ± 0.147 at hop 1 to 0.943 ± 0.016/0.941 ± 0.017 at hop 3, saturating near 0.976 at hop 7. The sharp collapse in standard deviation (SD) beyond hop 3 ($\sigma < 0.02$) demonstrates stable, data-wide manifold curvature, substantially exceeding values reported for CONCH [42] by Xiong et al. [67].

B Mathematical Preliminaries and Complete Proofs

Appendix B.1 collects established results invoked in our proofs, and Appendix B.2 the rectifiability results we compare against; the remaining subsections state and prove our own.

B.1 Background: Riemannian Primitives on $\mathcal{S}^{d-1}$, OT, Curvature-Dimension, Flow Matching

The proofs below invoke the Riemannian primitives of the unit hypersphere, OT on $\mathcal{S}^{d-1}$ and the RFM construction, each at a separate point; the sphere's curvature-dimension bound only motivates the choice of source. The geodesic distance, exponential map, and logarithmic map on $\mathcal{S}^{d-1}$ supply



(a) Spectral analysis

(b) Tangent drift

Figure 5: Intrinsic geometry of histopathology VFM feature manifolds on BRCA. (a) Eigenvalue spectrum of the ℓ_2-normalized feature covariance for UNI and UNI2-h, showing high covariance effective rank ($R_{\text{eff}} = 252/265$; distinct from the Jacobian-spectrum cut k^*). (b) Tangent drift $\mathcal{D}$ vs. graph-hop distance, demonstrating monotonically increasing misalignment of local tangent spaces, which is direct evidence of a curved, non-flat manifold structure.

the notation used throughout:

$$d_g(x, y) = \arccos(\langle x, y \rangle), \tag{10}$$

$$\operatorname{Exp}_x(v) = \cos(\|v\|)\, x + \sin(\|v\|)\, \frac{v}{\|v\|}, \quad v \in T_x\mathcal{S}^{d-1}, \tag{11}$$

$$\operatorname{Log}_x(y) = \frac{\theta}{\sin\theta}(y - \cos\theta\, x), \quad \theta = d_g(x, y). \tag{12}$$

OT on $\mathcal{S}^{d-1}$ takes this geodesic distance as its cost. The 2-Wasserstein distance $W_2^2(p_0, p_1) = \inf_{\gamma \in \Gamma(p_0,p_1)} \int d_g(x, y)^2\, d\gamma$ admits the dynamic Benamou–Brenier formulation [4], which measures a transport plan by the kinetic action $\int_0^1 \mathbb{E}\|v_t\|_g^2\, dt$ of a velocity field carrying p_0 to p_1, and Proposition 2 bounds that action for the bridge. On $\mathcal{S}^{d-1}$, OT maps inherit smoothness from the Ma–Trudinger–Wang (MTW) condition [41, 44], which supplies the per-factor C^2 control of v_t^{det} that the kernel-regression bias estimate of Proposition 3(ii) needs, with the extension across the N coupled tokens carried by hypothesis (CT) instead. The sphere satisfies $\mathrm{CD}(d-2, \infty)$ [2], with $\mathrm{Ric} \geq (d-2)g$ implying W_2-contraction of the heat semigroup at rate $d-2$ [61]; this makes $p_0 = \mathrm{Uniform}(\mathcal{S}^{d-1})$ the canonical source, and Section 4.1 adopts it on the product manifold for that reason.

FM [1, 37, 39, 40] learns a velocity field via the CFM loss $\mathcal{L}_{\text{CFM}} = \mathbb{E}[\|v_\theta - u_t\|^2]$; for Riemannian flows on $\mathcal{S}^{d-1}$ [14], the SLERP conditional path $\psi_t(x_0|x_1) = \mathrm{SLERP}(x_0, x_1, t)$ and conditional velocity $u_t = \operatorname{Log}_{x_t}(x_1)/(1-t)$ are the standard choices. Definition 1 keeps both and displaces the path off the geodesic by tangent-Gaussian noise, so the conditional velocity bounded in Proposition 2 is the expression above evaluated at the perturbed x_t rather than at the SLERP point μ_t.

B.2 Rectifiability Theory

The Euclidean rectifiability theory of Hertrich et al. [21] provides a sufficient condition, which Theorem 1 does not use:

Theorem (Sufficient Rectifiability) [21, Thm. 14]. *If $P_{X_0|X_1=x_1}$ is absolutely continuous with smooth positive density, then (X_0, X_1) is rectifiable.* Here rectifiable carries the sense of Hertrich et al. [21, Def. 1], under which the ordinary differential equation $\dot{z}_t = v_t(z_t)$ admits a unique solution. The hypothesis is a condition on the coupling's conditional law of the source given the data endpoint.

Proposition (Disconnected-support obstruction) [21, Prop. 10]. *There exist couplings with zero rectification loss whose intermediate marginals have disconnected support and whose rectification fixed point is not the OT coupling, invalidating earlier rectification–OT equivalence claims.*

Both results are stated and proved in $\mathbb{R}^d$; Theorem 1 uses neither and is proved directly on $\mathcal{S}^{d-1}$ (Appendix B.3).

The bridge perturbation does not supply the premise of Hertrich et al. [21, Thm. 14]. That premise is a condition on $P_{X_0|X_1=x_1}$, and the bridge is pinned at both endpoints ($\sigma(0) = \sigma(1) = 0$), so it leaves $P_{X_0|X_1=x_1}$ untouched; under our independent coupling $\pi = p_0 \otimes p_1$ that conditional is already $\mathrm{Uniform}(\mathcal{S}^{d-1})$, so the premise holds for a reason unrelated to the bridge. Theorem 1 does not depend on it in any case. Lemma 2, which also makes the Bayes denominator positive, gives the smooth conditional density of (X_0, X_1) given X_t directly on $\mathcal{S}^{d-1}$, which is exactly part (ii). Part (ii) with Proposition 2 then gives part (iii) by Picard–Lindelöf, i.e. the unique solution of $\dot{x} = v_t(x)$ that Hertrich et al. [21, Def. 1] call rectifiability. The bridge instead supplies the intermediate-marginal property (Proposition 1), which Hertrich et al. [21, Prop. 10] show to be the binding obstruction; see Appendix B.5.

B.3 Proof of Theorem 1: Three Consequences of the Bridge Perturbation

Lemma 1 (Exponential map Jacobian)**.** *For $v \in T_x\mathcal{S}^{d-1}$ with $0 < \|v\| = r < \pi$ (i.e. $\mathrm{Exp}_x(v)$ outside the cut locus of x):* $\det(d\,\mathrm{Exp}_x)_v = (\sin r/r)^{d-2}$.

Proof. In geodesic polar coordinates the Riemannian volume element is $\sin^{d-2}(r)\, dr\, d\omega$ and the Euclidean tangent-space element is $r^{d-2}\, dr\, d\omega$ [15]; the ratio is $(\sin r/r)^{d-2}$. □

Lemma 2 (Full support of conditional distribution)**.** *For any $\sigma > 0$ and $\mu \in \mathcal{S}^{d-1}$, the distribution of $X = \mathrm{Exp}_\mu(\sigma\epsilon)$ with $\epsilon \sim \mathcal{N}(0, \Pi_\mu)$ (the rank-$(d-1)$ tangent-space Gaussian, embedded in $\mathbb{R}^d$) has positive density on $\mathcal{S}^{d-1} \setminus \{-\mu\}$.*

Proof. The Gaussian $\sigma\epsilon$ has full support on the column space $T_\mu\mathcal{S}^{d-1} \cong \mathbb{R}^{d-1}$ of Π_μ. The exponential map is a diffeomorphism on the open ball of radius π in $T_\mu\mathcal{S}^{d-1}$ onto $\mathcal{S}^{d-1} \setminus \{-\mu\}$. By Lemma 1 and the change-of-variables formula, the principal-branch pushforward density is:

$$p_X(y) = \frac{1}{(2\pi\sigma^2)^{(d-1)/2}} \exp\left(-\frac{d_g(\mu, y)^2}{2\sigma^2}\right) \left(\frac{d_g(\mu, y)}{\sin(d_g(\mu, y))}\right)^{d-2}, \quad y \in \mathcal{S}^{d-1} \setminus \{-\mu\}.$$

Smoothness follows because $d_g(\mu, \cdot)^2$ is C^∞ on $\mathcal{S}^{d-1} \setminus \{-\mu\}$ and $r \mapsto (r/\sin r)^{d-2}$ is real-analytic on $[0, \pi)$. Positivity follows from positivity of the Gaussian factor. Wrap-around contributions through the cut locus are bounded by $P(\sigma\|\epsilon\| \geq \pi) \leq \exp(-\frac{1}{2}(\pi/\sigma - \sqrt{d-1})^2)$; in the deployed regime $\sigma_{\max} = 0.01 \ll \pi/\sqrt{d-1} \approx 0.098$ for $d = 1024$, this is below $e^{-3\times10^4}$ and does not affect positivity or smoothness. □

Full proof of Proposition 1. Fix $t \in (0, 1)$. Then $\sigma(t) = \sigma_{\max}\sin(\pi t) > 0$. By Lemma 2, $\mathrm{law}(X_t \mid x_0, x_1)$ has positive smooth density on $\mathcal{S}^{d-1} \setminus \{-\mu_t(x_0, x_1)\}$ for each (x_0, x_1) off the antipodal locus $\bar{\Delta}$.

Topological support. Let π be any coupling with $\pi(\bar{\Delta}) = 0$. For every nonempty open $O \subseteq \mathcal{S}^{d-1}$ and every $(x_0, x_1) \notin \bar{\Delta}$, the conditional density is strictly positive on O minus at most one point, so $P(X_t \in O \mid x_0, x_1) > 0$; integrating against π gives $P(X_t \in O) > 0$. Hence $\mathrm{supp}(\mathrm{law}(X_t)) = \mathcal{S}^{d-1}$. No absolute continuity of π is used, which matters because the rectification iterates of Corollary 1 are deterministic couplings.

Density. If in addition π is absolutely continuous, the map $(x_0, x_1) \mapsto -\mu_t(x_0, x_1)$ is a smooth submersion on $(\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}) \setminus \bar{\Delta}$, so for vol_g-a.e. $y \in \mathcal{S}^{d-1}$ the preimage $\{(x_0, x_1) : y = -\mu_t(x_0, x_1)\}$ is a codimension-$(d-1)$ submanifold and hence π-null. By Fubini, $\mathrm{law}(X_t) = \int \mathrm{law}(X_t \mid x_0, x_1)\, d\pi$ then has positive density vol_g-a.e. □

Proposition 2 (Bounded velocity and finite action)**.** *Under $\sigma(t) = \sigma_{\max}\sin(\pi t)$: $\|u_t(x_t \mid x_0, x_1)\|_g \leq \sigma_{\max}\pi\|\epsilon\| + \Omega$ for all $t \in (0, 1)$, where $\Omega = d_g(x_0, x_1)$.*

Proof. For $t \in [0,1)$, $\sin(\pi t) = \sin(\pi(1-t)) \le \pi(1-t)$ (using $\sin x \le x$ for $x \ge 0$), so $\sigma(t)/(1-t) \le \sigma_{\max}\pi$. The geodesic distance satisfies $d_g(x_t, \mu_t) \le \sigma(t)\|\epsilon\|$ unconditionally: when $\sigma(t)\|\epsilon\| < \pi$ (inside the injectivity ball), this holds with equality by the radial isometry Exp_{μ_t}; otherwise $d_g(x_t, \mu_t) \le \pi \le \sigma(t)\|\epsilon\|$ trivially since $\mathcal{S}^{d-1}$ has diameter π. The triangle inequality therefore yields, unconditionally:

$$\|u_t\|_g = \frac{d_g(x_t, x_1)}{1-t} \le \frac{d_g(x_t, \mu_t) + d_g(\mu_t, x_1)}{1-t} \le \frac{\sigma(t)\|\epsilon\| + (1-t)\Omega}{1-t} \le \sigma_{\max}\pi\|\epsilon\| + \Omega.$$

Using $\mathbb{E}[\|\epsilon\|^2] = d-1$, $\mathbb{E}\|\epsilon\| \le \sqrt{d-1}$ (Jensen), $\Omega \le \pi$, and independence of Ω and $\|\epsilon\|$, the action integral satisfies $\int_0^1 \mathbb{E}[\|v_t\|_g^2]\, dt \le (\sigma_{\max}\pi\sqrt{d-1} + \pi)^2 < \infty$. □

Proposition 3 (Approximate transport: an asymptotic rate in $\sigma_{\max}$)**.** *Let $\Omega_{\max}$ be an upper bound on the endpoint angle Ω of Eq. (2). Part (i) holds unconditionally. Parts (ii) and (iii) assume in addition hypothesis **(CT)**, that the deterministic marginal velocity v_t^{det} is C^2 on $(\mathcal{S}^{d-1})^N \times [\epsilon_t, 1-\epsilon_t]$ jointly in all N token coordinates, and hold in the limit $\sigma_{\max} \to 0$ at fixed $(N, d, \Omega_{\max}, \epsilon_t)$. The bridge-augmented flow satisfies:*

- *(i) boundary consistency, $\sigma(0) = \sigma(1) = 0$;*
- *(ii) velocity perturbation $\|\tilde{v}_t - v_t^{\mathrm{det}}\|_\infty = O(\sigma_{\max}^2 \sin^2(\pi t))$, uniformly on $[\epsilon_t, 1-\epsilon_t]$, where $\tilde{v}_t$ is the bridge-smoothed marginal velocity;*
- *(iii) $W_2\big((\Phi_1^{\tilde{v}})_{\#}p_0,\, p_1\big) = O(\sigma_{\max}^2)$, writing Φ_1^v for the time-1 flow map of a field v.*

Parts (ii) and (iii) are rates. The implied constants are uniform for $\sigma_{\max}$ in a bounded neighborhood of 0, depend on $(N, d, \Omega_{\max}, \epsilon_t)$, and are not quantified here; in particular Proposition 3 makes no finite-parameter guarantee at the deployed $\sigma_{\max} = 10^{-2}$ and $\epsilon_t = 10^{-5}$, and asserts no specific numerical transport error there.

Proof. (i) Immediate from $\sigma(0) = \sigma(1) = 0$.

(ii) The bridge-augmented velocity $\tilde{v}_t$ is a kernel-smoothed conditional expectation with bandwidth $\sigma(t)$. By on-manifold kernel-regression bias theory for closed Riemannian manifolds [49], the bias is $O(\sigma(t)^2)$ on $[\epsilon_t, 1-\epsilon_t]$ when v_t^{det} is C^2 (the Riemannian volume Jacobian of Exp_{μ_t} is $1 - \frac{d-2}{6}r^2 + O(r^4)$ for $r = \|\sigma(t)\epsilon\|$; the $O(r^2)$ term carries a dimensional $(d-2)/6$ factor that we absorb into the $O(\sigma(t)^2)$ bound, leaving leading-order bias unchanged). On a single $\mathcal{S}^{d-1}$, C^2 regularity of v_t^{det} follows from C^∞ regularity of the OT map (MTW theory [41, 44]) propagated through $v_t^{\mathrm{det}} = \mathrm{Log}_x(x_1)/(1-t)$, bounded on $[\epsilon_t, 1-\epsilon_t]$. On the product manifold $(\mathcal{S}^{d-1})^N$ this per-token argument does not suffice, because self-attention couples the tokens and MTW theory is a per-factor statement; the joint C^2 regularity is hypothesis (CT), and it is assumed, not proved. (CT) enters here and nowhere else: it bounds the transport rate of Proposition 3 and is used by neither Proposition 1 nor any part of Theorem 1. Note also that (CT) constrains v_t^{det}, the deterministic velocity, and not the bridge-smoothed marginal field v_t of Theorem 1(iii) (written $\tilde{v}_t$ here), whose regularity the smoothing itself produces.

(iii) On $[\epsilon_t, 1-\epsilon_t]$, the marginal velocity v_t is uniformly Lipschitz with constant $L = L(N, d, \Omega_{\max}, \sigma_{\max}, \epsilon_t)$. By Grönwall's inequality, $\sup_x d_g(\Phi_1^{\tilde{v}}(x), \Phi_1^{v^{\mathrm{det}}}(x)) \le C\sigma_{\max}^2$ with C depending on the same parameters; coupling $X \sim p_0$ through both flows gives a valid coupling of $(\Phi_1^{\tilde{v}})_{\#}p_0$ and $(\Phi_1^{v^{\mathrm{det}}})_{\#}p_0 = p_1$, so $W_2^2((\Phi_1^{\tilde{v}})_{\#}p_0, p_1) \le \mathbb{E}_{x\sim p_0}[d_g(\Phi_1^{\tilde{v}}(x), \Phi_1^{v^{\mathrm{det}}}(x))^2] \le C^2\sigma_{\max}^4$, giving $W_2 \le C\sigma_{\max}^2$. We note that C depends on ϵ_t via the velocity blow-up at $t \to 1$, and that the proof of (ii) absorbs the dimensional $(d-2)/6$ volume-Jacobian factor into the $O(\sigma(t)^2)$ bound. The $O(\sigma_{\max}^2)$ rate is what we claim; the absolute constant at the deployed cutoff $\epsilon_t = 10^{-5}$ is large, so Proposition 3 is not a numerical guarantee at the deployed $\sigma_{\max} = 0.01$. It does not, for instance, assert a 10^{-4} transport error. The qualitative smallness of the transport perturbation is corroborated empirically instead (Table 4, isotropic-decoder rows; Appendix C.1). □

Corollary 1 (Bridge resolution of the disconnected-support obstruction)**.** *Let π be any coupling on $\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$ giving no mass to the antipodal locus, $\pi(\bar{\Delta}) = 0$. This is a strictly weaker hypothesis than absolute continuity, and one that the deterministic couplings arising as rectification iterates satisfy. Then the bridge perturbation with $\sigma_{\max} > 0$ ensures $\mathrm{supp}(\mathrm{law}(X_t)) = \mathcal{S}^{d-1}$ for all $t \in (0,1)$, whatever the structure of p_1. Consequently, no rectification iterate can become a non-optimal fixed point of the kind Hertrich et al. [21, Prop. 10] identify, exhibiting the disconnected-support obstruction.*

Proof. By Lemma 2 the conditional density is strictly positive on $\mathcal{S}^{d-1} \setminus \{-\mu_t(x_0, x_1)\}$ for every pair $(x_0, x_1) \notin \bar{\Delta}$, so every nonempty open set has positive probability under any π with $\pi(\bar{\Delta}) = 0$; no absolute continuity of π is used. Hence $\mathrm{law}(X_t)$ has full topological support for every $t \in (0, 1)$ and every such π, including the deterministic rectification iterates. Hence no fixed point of the bridge–rectification map exhibits the disconnected-support pathology underlying the counterexample of Hertrich et al. [21]. □

Full proof of Theorem 1.

(i) Proposition 1.

(ii) By Bayes' rule (with respect to the Riemannian volume measure on $\mathcal{S}^{d-1}$ as the dominating reference): $p(x_0, x_1 \mid X_t = x) \propto p(X_t = x \mid x_0, x_1)\, d\pi$, where the conditional density $p(X_t = x \mid x_0, x_1)$ is given by Lemma 2 on the cut-locus complement (a full-measure subset of $\mathcal{S}^{d-1}$ for $d \geq 3$). The numerator is smooth on this domain (Gaussian pushforward through the exponential map). The denominator is positive (Lemma 2) and smooth on the cut-locus complement, using the per-direction Gaussian decay of Lemma 2 as the integrable dominating function for differentiation under the integral. On $(\mathcal{S}^{d-1})^N$ the conditional density of X_t given (x_0, x_1) is the product of the N per-token densities of Lemma 2, so it is positive and smooth exactly when each factor is, and the Bayes argument above applies verbatim with Π_{μ_t} replaced by $\bigoplus_n \Pi_{\mu_t^{(n)}}$. What is independent is the bridge noise, which is exact by construction.

(iii) On $[\epsilon_t, 1 - \epsilon_t]$: $v_t(x) = \mathbb{E}[u_t \mid X_t = x]$ is smooth by differentiation under the integral (using the Gaussian dominating function from (ii) and the bounded u_t from Proposition 2). Differentiating once more, $\|\nabla_x v_t\|_\infty \leq L$ with $L = L(d, \Omega_{\max}, \sigma_{\max}, \epsilon_t)$ finite on the closed interval (the kernel-derivative bound scales as $\sigma(t)^{-1}$, which is bounded below by $\sigma_{\max} \sin(\pi\epsilon_t) > 0$). The Picard–Lindelöf theorem for time-dependent Lipschitz vector fields on a compact Riemannian manifold [33] then gives existence and uniqueness of the initial value problem $\dot{x} = v_t(x), x(t_0) = x_0$ on $[\epsilon_t, 1 - \epsilon_t]$ for any t_0 in the same interval. Part (iii) also lifts to $(\mathcal{S}^{d-1})^N$: the argument uses only the dominating function from (ii), the velocity bound of Proposition 2 and $\sigma(t) \geq \sigma_{\max} \sin(\pi\epsilon_t) > 0$, all of which hold on the product, with N entering the Lipschitz constant L and not the hypotheses. In particular the cross-token hypothesis (CT) of Proposition 3 is not used anywhere in Theorem 1. □

B.4 Exact Conditional Velocity of the Stochastic Bridge

In FM the conditional velocity is the time derivative of the conditional path, and that is the sense in which Section 4.2 and Appendix C.2 use the term. For the bridge it is $\dot{X}_t$ under the conditioning $z = (x_0, x_1, \epsilon)$; the marginal field that transports the bridge's own marginals is $\mathbb{E}[\dot{X}_t \mid X_t = x]$, which admits no closed form. We give $\dot{X}_t$ in closed form.

Definition 1 fixes the marginal law of X_t at each t but not a joint process across t, so $\dot{X}_t$ requires choosing a coupling of the noise across time. We take parallel transport of ϵ along the SLERP curve μ. Parallel transport is a linear isometry, so every marginal in Definition 1 is unchanged and $t \mapsto X_t$ becomes differentiable.

Proposition 4 (Exact conditional velocity of the bridge)**.** *Write $r_t = \sigma(t)\|\epsilon\|$ and $w_t = P_{\mu_0 \to \mu_t}\epsilon / \|\epsilon\|$, so that $X_t = \cos(r_t)\, \mu_t + \sin(r_t)\, w_t$, and set $a = \langle \dot{\mu}_t, w_t \rangle$. Then*

$$\dot{X}_t = \big[\sigma'(t)\|\epsilon\| + a\big]\, \nu_t + \cos\big(\sigma(t)\|\epsilon\|\big)\, \dot{\mu}_t^{\perp}, \qquad \begin{aligned} \nu_t &= \cos(r_t)\, w_t - \sin(r_t)\, \mu_t, \\ \dot{\mu}_t^{\perp} &= \dot{\mu}_t - a\, w_t. \end{aligned} \tag{13}$$

The right-hand side is tangent to $\mathcal{S}^{d-1}$ at X_t and depends only on $(t, \mu_t, \dot{\mu}_t, X_t, \epsilon)$. Here $\nu_t = -\operatorname{Log}_{X_t}(\mu_t)/r_t$ is the outward radial unit vector at X_t of the geodesic emanating from μ_t. Setting $\sigma \equiv 0$ returns $\dot{X}_t = \dot{\mu}_t$, the deterministic SLERP velocity.

Proof. Parallel transport along μ keeps the ambient derivative of w_t normal to the sphere, so $\dot{w}_t = \lambda \mu_t$ for some scalar λ; differentiating $\langle w_t, \mu_t \rangle = 0$ gives $\lambda = \langle \dot{w}_t, \mu_t \rangle = -\langle w_t, \dot{\mu}_t \rangle = -a$, i.e. $\dot{w}_t = -a\, \mu_t$. Differentiating $X_t = \cos(r_t)\mu_t + \sin(r_t)w_t$ and using $\dot{r}_t = \sigma'(t)\|\epsilon\|$,

$$\dot{X}_t = \dot{r}_t\big(\cos(r_t)w_t - \sin(r_t)\mu_t\big) + \cos(r_t)\dot{\mu}_t - a\sin(r_t)\mu_t.$$

Decompose $\dot{\mu}_t = \dot{\mu}_t^{\perp} + a\, w_t$. The term $a\cos(r_t)w_t$ combines with $-a\sin(r_t)\mu_t$ into $a\,\nu_t$, giving Eq. (13). Tangency: $\langle X_t, \nu_t\rangle = -\cos r_t \sin r_t + \sin r_t \cos r_t = 0$ and $\langle X_t, \dot{\mu}_t^{\perp}\rangle = \sin(r_t)(\langle w_t, \dot{\mu}_t\rangle - a) = 0$. □

What the perturbation changes. Against the deterministic $\dot{\mu}_t$, the bridge alters the conditional velocity in exactly two ways. It adds a radial component along ν_t at rate $\sigma'(t)\|\epsilon\|$, and it attenuates the transporting drift $\dot{\mu}_t^{\perp}$ by $\cos r_t$, which is the basepoint derivative of Exp and hence a curvature effect. At the deployed values ($d = 1024$, $\sigma_{\max} = 10^{-2}$, $\|\epsilon\| \approx \sqrt{d-1}$) the attenuation $1 - \cos r_t$ is 5.1% at $t = 0.5$. The a term rides on ν_t alongside the radial rate and is small in root-mean-square, $\Omega/\sqrt{d-1} \approx 0.05$ with $\Omega = d_g(x_0, x_1)$.

Consistency with the inference velocity. The inference velocity is not intended to reproduce $\dot{X}_t$, and does not. The bridge acts only at training time: it is a $\sigma(t)$-bandwidth smoothing of the endpoint regression, not a path to be integrated. Algorithm 1 fits a normalized endpoint $\hat{x}_1$ under the chord loss Eq. (5), so the learned field is $\mathrm{Log}_x(\hat{x}_1)/(1-t)$ with $\hat{x}_1$ the normalized $\mathbb{E}[x_1 \mid X_t = x]$. Only the transport of p_0 to p_1 has to agree, and it agrees approximately rather than exactly. Proposition 3 quantifies the gap for $\tilde{v}_t = \mathbb{E}[u_t \mid X_t]$ rather than for the chord field, with part (ii) bounding $\tilde{v}_t - v_t^{\mathrm{det}}$ and part (iii) converting it into $W_2 = O(\sigma_{\max}^2)$, subject to the scoping stated there.

Why we regress the endpoint target rather than $\dot{X}_t$. CFM on the bridge as a conditional path would regress $\dot{X}_t$; we regress the SLERP-type endpoint target instead, and Proposition 3 bounds what that deliberate misspecification costs. The trade is safe because the radial channel nets to zero over the path, $\int_0^1 \sigma'(t)\,dt = \sigma(1) - \sigma(0) = 0$, so it transports nothing even though it is far from pointwise small. At the deployed values it peaks at $\sigma_{\max}\pi\sqrt{d-1} \approx 1.00$ at the endpoints, against a geodesic speed $\Omega \approx \pi/2$. Regressing it would add real variance for zero net transport, and the smoothing bias left behind is Proposition 3(ii)'s $O(\sigma_{\max}^2)$.

B.5 Support of the Intermediate Marginals Under Deterministic SLERP

The source marginal $p_0 = \mathrm{law}(X_0)$ and the intermediate marginals $\mathrm{law}(X_t)$ for $t \in (0,1)$ are constrained by different parts of the construction. p_0 constrains the source alone, and under deterministic SLERP the intermediate marginals are fixed by p_1 and t. Rectifiability in the sense of Hertrich et al. [21, Def. 1] asks for a unique solution of $\dot{z}_t = v_t(z_t)$, a property of the velocity field, and v_t follows from the conditional law of the endpoints given X_t rather than from p_0. A uniform source therefore does not supply it.

The unreachable cap. Since $d_g(\mathrm{SLERP}(x_0, x_1, t), x_1) = (1-t)\Omega$ with $\Omega = d_g(x_0, x_1) \le \pi$, for fixed x_1 the support of $\mathrm{law}(X_t \mid x_1)$ under deterministic SLERP is the closed ball $\bar{B}(x_1, (1-t)\pi)$, leaving the open cap $B(-x_1, t\pi)$ unreachable. With $p_1 = \delta_{x_1}$, $p_0 = \mathrm{Uniform}(\mathcal{S}^{d-1})$ and $t = \frac{1}{2}$, the support is the closed hemisphere centered at x_1 and the open hemisphere around $-x_1$ carries zero mass, half the sphere under a uniform source. In general

$$\mathrm{supp}\big(\mathrm{law}(X_t)\big) \;=\; \big\{y : d_g(y, \mathrm{supp}\, p_1) \le (1-t)\pi\big\},$$

so if p_1 lies in a cap of angular radius α about c, then a cap of radius $t\pi - \alpha$ about $-c$ is missed for every $t > \alpha/\pi$. Any $\sigma_{\max} > 0$ fills it, since Lemma 2's tangent Gaussian is positive on $\mathcal{S}^{d-1}\backslash\{-\mu_t\}$ and Proposition 1 then gives full support at every $t \in (0,1)$.

Consequences at $\sigma_{\max} = 0$. Off the support, the density of $\mathrm{law}(X_t)$ vanishes, so Bayes' rule does not define the conditional law of (X_0, X_1) given $X_t = x$ there, and the velocity it produces covers only a proper closed subset of $\mathcal{S}^{d-1}$. On the support, deterministic SLERP still misses the density property of Theorem 1(ii), because conditioning on $X_t = x$ leaves a direction ($d-2$ parameters) and a length (1), so the fiber is $(d-1)$-dimensional inside a $2(d-1)$-dimensional product and is singular with respect to the product Riemannian volume measure on $\mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$, admitting no density against it. The vanishing density off the support and the singular fiber on it are properties of the Bayes construction at $\sigma_{\max} = 0$. The deterministic velocity v_t^{det} may still be smooth and admit a unique solution of $\dot{z}_t = v_t^{\mathrm{det}}(z_t)$ for a given p_1.

C Additional Technical Details

C.1 Noise Schedule: Derivation and Ablation

KL derivation. The standard Brownian bridge B on $[0,1]$ has covariance $\mathrm{Cov}(B_s, B_t) = \min(s,t) - st$. The associated eigenproblem

$$\int_0^1 \big(\min(s,t) - st\big)\, \phi_n(s)\, ds \;=\; \lambda_n\, \phi_n(t)$$

is solved by differentiating twice, giving $\phi_n'' = -\lambda_n^{-1}\phi_n$ with $\phi_n(0) = \phi_n(1) = 0$, hence

$$\phi_n(t) = \sqrt{2}\,\sin(n\pi t), \qquad \lambda_n = \frac{1}{(n\pi)^2}, \qquad n \geq 1,$$

and the KL expansion $B_t = \sum_{n\geq 1} Z_n \sqrt{2}\,\sin(n\pi t)/(n\pi)$ with $Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, whose variance $\sum_{n\geq 1} 2\sin^2(n\pi t)/(n\pi)^2$ sums to $t(1-t)$ and so recovers the bridge covariance. Mode n contributes a marginal SD proportional to $|\sin(n\pi t)|/(n\pi)$.

We retain the $n{=}1$ mode alone. It is the unique KL mode strictly positive on the open interval $(0,1)$: every $n \geq 2$ mode has $n-1$ interior zeros, at which the perturbation would vanish and the bridge would degenerate to the deterministic path at interior times, defeating the purpose of the construction. Rescaling the retained mode so that its peak is $\sigma_{\max}$ gives $\sigma(t) = \sigma_{\max}\sin(\pi t)$, with total noise budget $\int_0^1 \sigma(t)^2\, dt = \sigma_{\max}^2/2$.

Concentration ceiling on $\sigma_{\max}$. The perturbation must stay inside the injectivity radius of $\mathcal{S}^{d-1}$. With $\epsilon \sim \mathcal{N}(0, \Pi_{\mu_t})$ one has $\|\epsilon\| \approx \sqrt{d-1}$, so the typical geodesic displacement is $\sigma(t)\|\epsilon\| \leq \sigma_{\max}\sqrt{d-1}$, and requiring this to stay below π gives $\sigma_{\max} \lesssim \pi/\sqrt{d-1} \approx 0.098$ at $d = 1024$. This ceiling comes from the analysis rather than from search, so the admissible window is fixed by d; the largest value we test, 2×10^{-2}, sits well below the ceiling of 0.098, and where the optimum sits inside the window remains empirical.

Why a bridge-type perturbation rather than variance-exploding (VE) or variance-preserving (VP) noise.

(i) *Exact boundary conditions*: $\sigma(0) = \sigma(1) = 0$ ensures $X_0 = x_0$ and $X_1 = x_1$ exactly; VE-type processes do not naturally satisfy $\sigma(1) = 0$.

(ii) *Intrinsic to $\mathcal{S}^{d-1}$*: VP-type linear interpolation $(1-\beta)x_0 + \beta x_1 + \sigma\epsilon$ leaves the sphere; our construction uses the exponential map, keeping all intermediate points on $\mathcal{S}^{d-1}$.

(iii) *Controlled noise budget*: $\int_0^1 \sigma(t)^2\, dt = \sigma_{\max}^2/2$ directly controls the transport rate via Proposition 3.

Empirical comparison of schedules. We put the deployed kl1 schedule against three alternatives on BRCA. All four share the same SLERP backbone, the tangent Gaussian and the exponential map; only the scalar amplitude $\sigma(t)$ changes. kl1 and kl2 truncate the KL expansion above at one and two modes; under the common normalization both peak at $\sigma_{\max}$, so the pair tests first-mode sufficiency. si is the stochastic-interpolant noise profile $\gamma(t) = \sqrt{2t(1-t)}$ [1], rescaled so its peak is also $\sigma_{\max}$, making the comparison one of shape rather than amplitude. ve is an unpinned monotone control: $\sigma_{\max}\sqrt{t}$ is the marginal SD of the constant-diffusion member of the VE family [55], and is a representative unpinned profile rather than the schedule VE implementations usually deploy.

Noise-budget accounting. At a matched peak, $\int_0^1 \sigma^2$ is the total noise each schedule spends, so the table separates shape from volume. Under a uniform reading of t the unpinned control spends exactly kl1's budget, so what it loses cannot be charged to a larger budget; training draws t from a logit-normal law centered below $1/2$, under which ve in fact spends about 38% less than kl1, and it still loses 3.29 gFID against removing the bridge altogether (Table 12, top row). Spending less noise and placing it at the data end is worse than spending none, which is what pinning at $\sigma(1) = 0$ buys. ve is itself pinned at the source, $\sigma(0) = 0$, so the comparison bears on the data end alone. si shares

Table 11: Stage-1 noise-schedule ablation on BRCA (UNI encoder), full 90k decoder schedule, single checkpoint. $\int_0^1 \sigma^2$ is in units of $\sigma_{\max}^2$.

Schedule	$\sigma(t)$	$\sigma(1)$	$\int_0^1 \sigma^2$	rFID ↓	gFID ↓
kl1 (KL mode n=1; deployed)	$\sigma_{\max} \sin(\pi t)$	0	$\frac{1}{2}$	**2.23**	6.16
kl2 (KL modes n=1, 2)	$\sigma_{\max} \sqrt{\sin^2 \pi t + \frac{1}{4} \sin^2 2\pi t}$	0	$\frac{5}{8}$	3.33	**5.64**
si (stochastic interpolant)	$\sigma_{\max} \cdot 2\sqrt{t(1-t)}$	0	$\frac{2}{3}$	3.28	7.15
ve (VE, unpinned)	$\sigma_{\max} \sqrt{t}$	$\sigma_{\max}$	$\frac{1}{2}$	4.65	10.58

kl1's pinning and peak but spends 33% more on the uniform reading and loses on both metrics. kl2 does attain a better gFID, on a 25% larger budget and a 49% worse rFID (3.33 against 2.23); we retain kl1 because it is the schedule the theory of Section 4.2 is stated for and the configuration every other result in this paper is measured on.

Bridge noise scale $\sigma_{\max}$. Table 12 sweeps the bridge amplitude on BRCA. At $\sigma_{\max} = 0$ the construction reverts to deterministic SLERP and records the worst rFID of the four settings, 4.56. Raising the amplitude to 5×10^{-3} improves both metrics, and the deployed 10^{-2} gives the best rFID, 2.23, and the best gFID, 6.16. Doubling to 2×10^{-2} reverses both, costing 1.46 rFID and 1.72 gFID against the deployed value and recording the worst gFID of the four, 7.88. The deployed value therefore sits in the interior of the sweep on both metrics.

Table 12: Stage-1 bridge noise scale $\sigma_{\max}$ on BRCA, full 90k decoder schedule, single checkpoint.

$\sigma_{\max}$	rFID ↓	gFID ↓
0 (no bridge)	4.56	7.29
5×10^{-3}	3.22	6.93
$\mathbf{10^{-2}}$ *(deployed)*	**2.23**	**6.16**
2×10^{-2}	3.69	7.88

C.2 x-Prediction Formulation and Chord-vs-Geodesic Loss

The conditional velocity contains the factor $\sigma'(t)/\sigma(t) = \pi \cot(\pi t)$, which diverges at $t = 0$ and $t = 1$. The velocity itself stays finite, because $\pi \cot(\pi t)$ multiplies the radial displacement $d_g(\mu_t, X_t) = \sigma(t)\|\epsilon\|$, which vanishes at the same rate (Proposition 4). The parameterization is ill-conditioned at the endpoints, not the field. The x-prediction formulation [36] avoids this: the network $f_\theta(x_t, t)$ predicts the data endpoint with normalized output $\hat{x}_1 = f_\theta(x_t, t)/\|f_\theta(x_t, t)\|$, and at inference the velocity is recovered as $v_\theta(x_t, t) = \mathrm{Log}_{x_t}(\hat{x}_1)/(1-t)$. The $\sin(\pi t)/(1-t) \to \pi$ cancellation (Proposition 2) keeps the velocity target $u_t = \mathrm{Log}_{x_t}(x_1)/(1-t)$ bounded as $t \to 1$. Time is sampled per sample from the logit-normal distribution of Esser et al. [18], $t = \mathrm{sigmoid}(u)$ with $u \sim \mathcal{N}(-0.5, 1)$, clamped to $[\epsilon_t, 1-\epsilon_t]$ with $\epsilon_t = 10^{-5}$; no resolution-dependent time shift is applied, neither in training nor on the uniform sampling grid.

Chord vs. geodesic loss (full derivation). The natural Riemannian (geodesic) training loss measures velocity error in the tangent space via the logarithmic map:

$$\mathcal{L}_{\text{geo}}(\theta) = \mathbb{E}_{t,\,(x_0,x_1)\sim\pi,\,\epsilon}\left[\frac{\|\,\mathrm{Log}_{x_t}(\hat{x}_1) - \mathrm{Log}_{x_t}(x_1)\|_g^2}{(1-t)^2}\right]. \tag{14}$$

However, the Log map involves arccos (cf. Eq. (12)), whose Jacobian diverges as $\theta = \arccos(\langle x_t, \hat{x}_1\rangle) \to 0$, i.e. when the prediction collapses toward the current point x_t, a regime that becomes unavoidable near $t = 1$ where $x_t \approx x_1$. This creates a poorly conditioned gradient landscape. The chord loss Eq. (5) replaces the tangent-space distance with the ambient ℓ_2 distance between normalized predictions and shares the same per-sample minimizer ($\hat{x}_1 = x_1$) and the same inference velocity $v_\theta(x_t, t) = \mathrm{Log}_{x_t}(\hat{x}_1)/(1-t)$; the training gradient landscape changes, avoiding the arccos singularity.

Algorithm 1: STREAM Stage 1 DiT Training

Input: Precomputed data features $\{\mathbf{Z}_{1,j}\}_{j=1}^{B} \sim p_{\text{data}}$ on $(\mathcal{S}^{d-1})^N$, noise level $\sigma_{\max}$

for *each minibatch* **do**
  Sample $\mathbf{Z}_{0,j} \sim \text{Uniform}((\mathcal{S}^{d-1})^N)$ for $j = 1, \ldots, B$; `// independent coupling`
  Sample $t_j = \text{sigmoid}(u_j)$, $u_j \sim \mathcal{N}(-0.5, 1)$, clamp to $[\epsilon_t, 1-\epsilon_t]$, for $j = 1, \ldots, B$;
  for *each sample* j *and token* $n = 1, \ldots, N$ **do**
    $\mu_{t,j}^{(n)} \leftarrow \text{SLERP}(z_{0,j}^{(n)}, z_{1,j}^{(n)}, t_j)$;
    Sample $\epsilon_j^{(n)} \sim \mathcal{N}(0, \Pi_{\mu_{t,j}^{(n)}})$;
    $z_{t,j}^{(n)} \leftarrow \text{Exp}_{\mu_{t,j}^{(n)}}(\sigma_{\max} \sin(\pi t_j) \cdot \epsilon_j^{(n)})$;
  end
  Predict $\hat{\mathbf{Z}}_{1,j} = f_\theta(\mathbf{Z}_{t,j}, t_j)$; normalize each token: $\hat{z}_{1,j}^{(n)} \leftarrow \hat{z}_{1,j}^{(n)} / \|\hat{z}_{1,j}^{(n)}\|$;
  $\mathcal{L} \leftarrow \frac{1}{BN} \sum_{j=1}^{B} \sum_{n=1}^{N} \min\big(\|\hat{z}_{1,j}^{(n)} - z_{1,j}^{(n)}\|^2 / \max(1-t_j, 10^{-3})^2,\ 10^4\big)$; `// chord loss`
  Update θ by AdamW on $\mathcal{L}$ with gradient-norm clipping at 5.0; update the EMA of θ;
end

C.3 Stage 1 and Stage 2 Training Algorithms

Algorithm 1 summarizes Stage 1 training of the LightningDiT-XL generator f_θ on bridge-perturbed VFM embeddings (Definition 1) with the chord loss Eq. (5). For each minibatch, source latents $\mathbf{Z}_0$ are sampled from the uniform distribution on $(\mathcal{S}^{d-1})^N$, paired with precomputed data latents $\mathbf{Z}_1$ from the encoder; each sample j draws its own time t_j; per-token bridge interpolants $z_{t,j}^{(n)} = \text{Exp}_{\mu_{t,j}^{(n)}}(\sigma_{\max} \sin(\pi t_j)\, \epsilon_j^{(n)})$ are formed with $\epsilon_j^{(n)} \sim \mathcal{N}(0, \Pi_{\mu_{t,j}^{(n)}})$; and the network's normalized prediction is regressed against the data endpoint with the $1/(1-t)^2$ weighting absorbed into the chord loss. The deployed loss floors $1 - t_j$ at 10^{-3} and caps each per-token term at 10^4.

Algorithm 2 summarizes Stage 2 decoder training. The trained DiT enters only through the precomputed cache of centroids $\bar{\mathbf{Z}}_c$ and per-(centroid, token) bases U_H (Appendix D). Each minibatch loads precomputed features $\mathbf{Z}_j = E(I_j)$, assigns each sample to one centroid, perturbs each token with anisotropic noise from that centroid's basis for the token, projects back onto $\mathcal{S}^{d-1}$, and decodes; the 4-component loss Eq. (7) combines pixel-space ℓ_1 (a per-pixel mean), LPIPS, a token-averaged cosine round-trip term targeting the clean latent (forcing the decoder to denoise), and an adaptive adversarial term that is active only in phase C of Table 14.

The cosine round-trip loss in Stage 2 targets the clean features z, not the noisy input $\tilde{z}$. Per token, the noise is $n = \Pi_z(\sigma_H U_H U_H^\top \eta + \sigma_L (I_d - U_H U_H^\top)\eta)$, $\eta \sim \mathcal{N}(0, I_d)$, with U_H from the assigned centroid; at $z = \bar{z}_c$ its covariance is exactly Eq. (6). As $n \in T_z\mathcal{S}^{d-1}$, $\tilde{z} = (z+n)/\|z+n\|$ agrees with $\text{Exp}_z(n)$ up to $O(\|n\|^3)$. Adversarial adaptive weight: $\lambda_{\text{adapt}} = \|\nabla_{\text{last}} \mathcal{L}_{\text{rec}}\| / (\|\nabla_{\text{last}} \mathcal{L}_{\text{adv}}\| + 10^{-6})$, clamped to $[0.05, 10^4]$, where $\mathcal{L}_{\text{rec}}$ is the non-adversarial part of Eq. (7) (ℓ_1, LPIPS and cosine round-trip) and ∇_{last} the gradient at the decoder's last layer.

C.4 Integration Approximation and Generation Algorithm

The deployed midpoint sampler (Algorithm 3) steps with the exact per-token exponential map, so it has no retraction error; the midpoint velocity is carried from $T_{z_{1/2}}\mathcal{S}^{d-1}$ to $T_z\mathcal{S}^{d-1}$ by the projection Π_z instead of parallel transport, which changes the step only at $O(dt^3)$. Every operation acts per token, since a global operation on $\mathcal{S}^{Nd-1}$ would destroy the product structure. The code floors $1-t$ at 10^{-3}, which is inactive on this grid. The scheme is second-order Runge–Kutta, with $O(dt^3)$ per-step error (global $O(dt^2)$) against $O(dt^2)$ ($O(dt)$ global) for a Riemannian Euler step, so fewer steps suffice for matched accuracy.

Algorithm 2: STREAM Stage 2 Decoder Training

Input: Frozen encoder E, centroids $\{\bar{\mathbf{Z}}_c\}_{c=1}^{K} \subset (\mathcal{S}^{d-1})^N$ and per-(centroid, token) bases $\{U_H(\bar{z}_c^{(n)})\}$ from the DiT's cache, noise scales σ_H, σ_L, loss weights $(\lambda_{\ell_1}, \lambda_{\text{LPIPS}}, \lambda_{\cos}, \omega_G)$, phase boundaries $s_B < s_C$

for *each step* s *and minibatch* $\{(I_j, \mathbf{Z}_j)\}_{j=1}^{B}$, $\mathbf{Z}_j = E(I_j)$ *precomputed* **do**
- Apply random horizontal and vertical flips (each w.p. $\frac{1}{2}$) and a rotation by $k \cdot 90°$, $k \sim \mathcal{U}\{0,1,2,3\}$, identically to I_j and to the 16×16 token grid of $\mathbf{Z}_j$;
- $c_j \leftarrow \arg\max_c \sum_{n=1}^{N} \langle z_j^{(n)}, \bar{z}_c^{(n)} \rangle$; `// nearest centroid, per sample`
- **for** *each token* $n = 1, \ldots, N$ **do**
  - $U \leftarrow U_H(\bar{z}_{c_j}^{(n)})$; sample $\eta \sim \mathcal{N}(0, I_d)$;
  - $n_j^{(n)} \leftarrow \Pi_{z_j^{(n)}}\left(\sigma_H\, UU^\top \eta + \sigma_L (I_d - UU^\top)\, \eta\right)$; `// tangent noise`
  - $\tilde{z}_j^{(n)} \leftarrow (z_j^{(n)} + n_j^{(n)}) / \|z_j^{(n)} + n_j^{(n)}\|$; `// retract to` $\mathcal{S}^{d-1}$
- **end**
- Decode: $\hat{I}_j \leftarrow D_\phi(\tilde{\mathbf{Z}}_j)$; round-trip: $z_{\text{rt},j} \leftarrow E(\hat{I}_j)$; `// frozen encoder`
- $\mathcal{L}_{\text{rec}} \leftarrow \lambda_{\ell_1} \|\hat{I}_j - I_j\|_1 + \lambda_{\text{LPIPS}}\, \text{LPIPS}(\hat{I}_j, I_j) + \lambda_{\cos}\,(1 - \langle z_{\text{rt},j}, z_j \rangle)$;
- $\mathcal{L}_{\text{dec}} \leftarrow \mathcal{L}_{\text{rec}} + \omega_G\, \lambda_{\text{adapt}}\, \mathcal{L}_{\text{adv}}$ if $s \geq s_C$, else $\mathcal{L}_{\text{dec}} \leftarrow \mathcal{L}_{\text{rec}}$;
- Update ϕ by AdamW on $\mathcal{L}_{\text{dec}}$; update the EMA of ϕ;
- **if** $s \geq s_B$ **then** update the discriminator (hinge loss);

end

Algorithm 3: STREAM Generation

Input: Trained DiT f_θ (EMA weights), number of steps $n_{\text{steps}} = 25$, decoder D_ϕ

Sample $\mathbf{Z}_0 \sim \text{Uniform}((\mathcal{S}^{d-1})^N)$; $\mathbf{Z} \leftarrow \mathbf{Z}_0$;
$dt \leftarrow 1/n_{\text{steps}}$; `// uniform time grid`
for $i = 0, \ldots, n_{steps} - 1$ **do**
- $t \leftarrow i/n_{\text{steps}}$;
- $\hat{\mathbf{Z}}_1 \leftarrow f_\theta(\mathbf{Z}, t)$; normalize each token;
- **for** *each token* n **do**
  - $v^{(n)} \leftarrow \text{Log}_{z^{(n)}}(\hat{z}_1^{(n)})/(1-t)$; `// velocity at` t
  - $z_{1/2}^{(n)} \leftarrow \text{Exp}_{z^{(n)}}\left((dt/2) \cdot v^{(n)}\right)$; `// half step`
- **end**
- $\hat{\mathbf{Z}}_1^{(1/2)} \leftarrow f_\theta(\mathbf{Z}_{1/2}, t + dt/2)$; normalize each token ; `// second DiT call`
- **for** *each token* n **do**
  - $v_{\text{mid}}^{(n)} \leftarrow \text{Log}_{z_{1/2}^{(n)}}(\hat{z}_1^{(1/2),(n)})/(1-t-dt/2)$;
  - $z^{(n)} \leftarrow \text{Exp}_{z^{(n)}}\left(\Pi_{z^{(n)}}(dt \cdot v_{\text{mid}}^{(n)})\right)$; `// full step from` $z^{(n)}$
- **end**

end
Normalize each token of $\mathbf{Z}$;
return $D_\phi(\mathbf{Z})$;

D STREAM Training Hyperparameters

Architectures. Table 13 gives the four components of the pipeline and their configurations; the training settings that drive them are in Table 14.

Fair-comparison protocol. Baselines vary in batch size and reported step count. To equalize compute, we train every baseline and STREAM for the ImageNet-equivalent of 80 epochs, setting each method's number of optimization steps from its batch size and the dataset size. All hyperparameters other than schedule length follow each method's published configuration.

Table 13: Component architectures of the deployed STREAM pipeline.

Component	Backbone	Configuration
DiT generator	LightningDiT-XL [71]	28 layers, hidden 1152, 16 heads, SwiGLU feed-forward network, RMSNorm, rotary positional embeddings, AdaLN-Zero modulation (time-only), $\sim$ 676M parameters; per-token ℓ_2-normalized output for x-prediction on $\mathcal{S}^{d-1}$
Decoder	ViT-XL decoder of RAE [77]	Tanh output mapped to $[-1, 1]$
Encoder	UNI ViT-L/16 [13]	Frozen; 256 ℓ_2-normalized tokens of dimension 1024 from 256$\times$ 256 patches
Discriminator	Kaiko ViT-S/8 [29]	A histopathology-specific ViT pretrained on large-scale tissue patches, with frozen backbone and trainable spectrally-normalized projection heads, so the adversarial signal is computed in a histopathology-aware feature space rather than a natural-image one

Table 14: STREAM training hyperparameters.

Setting	Stage 1 (DiT)	Stage 2 (Decoder)
Architecture	LightningDiT-XL	ViT-XL decoder + Kaiko disc.
Parameters	$\sim$676M	ViT-XL + frozen UNI + frozen Kaiko
Loss	Chord loss Eq. (5)	ℓ_1 + LPIPS + cos round-trip + GAN
Loss weights	—	λ_{ℓ_1}=1.0, λ_{LPIPS}=0.5, $\lambda_{\cos}$=0.5, ω_G=0.75
Time sampling	Logit-Normal$(-0.5, 1)$ on $[\epsilon_t, 1-\epsilon_t]$, ϵ_t=10^{-5}	—
Bridge $\sigma_{\max}$	0.01	—
Noise injection	—	Anisotropic SVD (Eq. (6)): σ_H=0.002, σ_L=0.02, τ=0.90
Spectral basis	—	K=50 centroids, m=600 tangent probes, time anchor t=0.99, deployed k^*=236, $\dim U_L = d - 1 - k^* = 787$
Optimizer	AdamW, β=$(0.9, 0.95)$	AdamW, β=$(0.9, 0.95)$, weight decay 0.05
Learning rate	4×10^{-4}, constant	$2\times10^{-4} \to 2\times10^{-5}$ cosine
Warmup	5k steps	3k steps
Schedule	—	3-phase: A (0–17.5%) recon; B (17.5–20%) disc; C (20–100%) full
Steps	95k	90k
Batch size	1024	512
Grad clip	5.0	—
Exponential moving average decay	0.9999	0.9978
Generation	25 midpoint Euler steps	—

SVD of the velocity-field Jacobian (computation). The cached basis is extracted at a single time anchor, $t = 0.99$: one basis per (centroid, token), neither time-averaged nor recomputed during sampling. The $1/(1-t)$ prefactor of Section 4.2 is a positive scalar, so it rescales every singular value equally and moves no singular vector, no energy ratio and no k^*; anchoring near t=1 is therefore not by itself a blow-up artifact. The single-anchor cache is an approximation, and Appendix H.2 reports how much the basis moves with the anchor.

Extraction uses `torch.func.jvp` for batched forward-mode AD on the DiT's EMA weights, with $v_\theta = \text{Log}_{\mathbf{Z}}(\hat{\mathbf{Z}}_1)/(1-t)$ at $t = 0.99$. For centroid c and token n, $m = 600$ Gaussian probes (seed $cN + n$) are projected onto $T_{\bar{z}_c^{(n)}}\mathcal{S}^{d-1}$ and unit-normalized, not orthogonalized, and perturb token n only; the token-n block of the m Jacobian-vector products (JVPs), projected onto the same tangent space, forms a $d \times m$ response matrix whose exact thin SVD gives the cached vectors and spectrum s_i. Because forward-mode JVPs apply J rather than $J^\top$, the returned vectors are left singular vectors spanning $\text{range}(J)$. Appendix G.3 retrains the decoder on the right singular vectors instead and finds the left basis better by 1.15 rFID and 1.25 gFID.

One $k^* = 236$ serves every cell: the $\tau = 0.90$ cutoff of the element-wise median of the 50×256 cached spectra. U_L is the orthogonal complement of U_H in $T_{\bar{z}_c^{(n)}}\mathcal{S}^{d-1}$, of dimension $d - 1 - k^* = 787$. For the centroids, 50k random training features (seed 42) are mean-pooled over tokens, ℓ_2-normalized and clustered into $K = 50$ groups by mini-batch k-means; $\bar{\mathbf{Z}}_c \in (\mathcal{S}^{d-1})^N$ is its group's per-token mean, renormalized per token. In decoder training each sample goes to the centroid maximizing $\sum_n \langle z^{(n)}, \bar{z}_c^{(n)} \rangle$ on its full token grid (Algorithm 2).

D.1 Computational Cost

All figures for the training-time computational cost for STREAM (Table 15) are measured from our training logs and file sizes for the deployed STREAM configuration on one node of $8\times$ NVIDIA H200 GPUs. The STREAM-specific cost is the JVP and SVD stage alone: 69 GPU-h, 12.7% of the pipeline. Everything else is what any standard latent generative model with a learned diffusion model and a pixel decoder at a similar size costs.

Table 15: Training-time cost of the deployed STREAM pipeline on BRCA.

Stage	Wall-clock	GPU-hours
Encoder feature extraction	5.2 h	42
DiT training, 95k steps	16.9 h	135
k-means (K=50), JVP probes, SVD, basis caching	8.6 h	69
Decoder training, 90k steps	37.4 h	299
Total	**$\sim$68.1 h**	**$\sim$545**

D.2 Inference: NFE and quality

The deployed midpoint sampler evaluates f_θ twice per step (Algorithm 3), so the 25 steps behind Table 3 (Table 14) correspond to NFE $= 50$, and RAE and SVG are run at that same NFE. All points below use 50k samples and three independent sampling runs (Appendix E.6) on BRCA with the UNI encoder. They are a separate re-measurement of the models in Table 3, which reports a single run at the same sample count, so the RAE and SVG values differ from it.

Table 16: gFID against NFE on BRCA with the UNI encoder, mean $\pm$ SD over 3 sampling seeds at 50k samples per point. The midpoint sampler spends two DiT calls per step, so NFE 6 and 26 stand in for NFE 5 and 25, and all models run at the same NFE.

NFE	STREAM	RAE	SVG
6	49.70 ± 0.09	55.03 ± 0.05	55.80 ± 0.07
10	27.05 ± 0.06	29.87 ± 0.06	30.53 ± 0.18
26	9.31 ± 0.02	18.27 ± 0.04	18.82 ± 0.15
50	6.12 ± 0.04	15.88 ± 0.04	17.52 ± 0.14

STREAM leads at every NFE (6, 10, 26 and 50), and at NFE $= 26$ it already beats both baselines at NFE $= 50$, 9.31 against 15.88 and 17.52 (Table 16). At the matched NFE $= 50$ it is also the cheapest of the three to sample, at 12.14 TFLOPs and 94.8 ms per image against 16.09 and 156.7 ms for RAE and 24.16 and 175.0 ms for SVG (Table 17). STREAM therefore reaches the better sample quality at the lower inference cost, and reaches the baselines' NFE $= 50$ quality on half their function evaluations.

Table 17: Inference cost per generated image at NFE $= 50$. FLOPs are counted over the complete sampling call (every network evaluation plus the decoder). Latency is wall-clock at a generation batch of 32, divided by the batch, on a single NVIDIA H200.

Model	TFLOPs / image	ms / image
STREAM	**12.14**	**94.8**
RAE	16.09	156.7
SVG	24.16	175.0

E Additional Experiments and Ablations

Every cell in this appendix uses UNI as the encoder on BRCA unless stated otherwise, so the encoder is held fixed and only geometry, bridge, decoder regularization and backbone move. Each table is internally budget-matched; comparisons are within-table throughout.

E.1 Additional experiment: all six methods retrained on an independent non-TCGA dataset

Table 18: All six methods retrained from scratch on SPIDER-skin, with the encoder held at UNI throughout.

SPIDER-skin, encoder = UNI	rFID ↓	gFID ↓
STREAM	4.65	**9.19**
ZoomLDM	7.85	11.14
REPA-E	17.76	18.00
RAE	7.12	16.52
SVG	**1.30**	22.24
PixCell	3.32	453.28

SPIDER-skin [46] is an independent non-TCGA skin dataset, and all six methods are retrained on it from scratch (Table 18). STREAM reaches the best gFID at 9.19, with ZoomLDM next at 11.14, while SVG and PixCell hold the better rFID and perform poorly on gFID. As in Table 3, STREAM gives the best generation quality at little cost to reconstruction.

E.2 Ablation on prediction geometry

We test whether the prediction is constrained to the manifold by training STREAM with identical design and hyperparameters, varying only the underlying geometry the model is trained on. In the Euclidean row the predicted endpoint is not ℓ_2-normalized before the loss, so the model is fit as a Euclidean regressor instead of on the sphere. Both cells run at the full 90k decoder schedule rather than the matched 45k budget of Table 4, so this pair is read within itself.

Table 19: Geometry control on BRCA. Both rows are STREAM; only the ℓ_2 renormalization of the prediction differs.

	rFID ↓	gFID ↓
Euclidean: prediction not projected onto the sphere	4.52	16.28
Riemannian: projected (deployed)	**2.23**	**6.16**

Euclidean geometry vastly underperforms the deployed Riemannian geometry, at 16.28 gFID against 6.16 (Table 19). However, geometry is necessary but not sufficient for state-of-the-art generation. The first row of Table 4, at the 45k decoder budget, is spherical FM with neither of our STREAM components, and it performs worse than the ZoomLDM baseline of Table 3. This control holds the model completely fixed and therefore excludes architectural confounds; the Euclidean grid of Appendix E.3 does not, and we keep the two claims separate for that reason.

E.3 Ablation on STREAM's components in a Euclidean pipeline

Table 20: Both components rebuilt RAE-style in a Euclidean latent space (rFID / gFID), BRCA, UNI encoder.

RAE	Isotropic	Anisotropic
No bridge	6.00 / 24.39	5.80 / 25.38
Bridge	6.00 / 23.89	5.82 / 24.06

Next, we ask whether STREAM's components of bridge perturbation and anisotropic decoder can be applied to our closest Euclidean counterpart, RAE. The same ablation as Table 4 performed on RAE shows that adding both components brings no benefit (Table 20), empirical evidence that STREAM's

design is specific to the sphere. These four cells use a plain LightningDiT-XL without the Decoupled Diffusion Transformer (DDT) head [64], so that latent dimensions match STREAM and the SVD extraction is clean. They are therefore not comparable to the paper-faithful RAE baseline of Table 3, which uses the DDT head.

E.4 Ablation on prediction target and loss

We ask which pairing of loss space and prediction target the sphere selects, crossing an x-space (chord) loss against a v-space (geodesic) loss and endpoint prediction against velocity prediction. On $\mathcal{S}^{d-1}$ the metric is the Euclidean inner product restricted to the tangent space, so a velocity loss admits no extrinsic-versus-intrinsic choice and the chord-versus-geodesic choice arises only once endpoints are regressed. RFM methods usually define their target in tangent space, whether the network emits a velocity [14] or an endpoint, and none regresses endpoints in the ambient chord metric, so running this comparison at our manifold dimension (N=256 spheres at d=1024, i.e. 261,888 tangent dimensions) is, to our knowledge, new.

Table 21: Stage-1 prediction/loss parameterization on BRCA, full 90k decoder schedule, single checkpoint.

Parameterization	rFID ↓	gFID ↓
x loss (chord) + x-prediction (ours)	**2.23**	**6.16**
Standard RFM: v loss (geodesic) + v-prediction	4.78	7.84
x loss (chord) + v-prediction	4.73	11.26
v loss (geodesic) + x-prediction	3.12	12.57

The deployed pairing beats standard RFM on both metrics, 2.23 against 4.78 rFID and 6.16 against 7.84 gFID, and the two mismatched cells give the worst gFID of the four combinations (Table 21).

E.5 Ablation on discriminator

We test whether the histopathology-specific Kaiko ViT-S/8 that supplies the decoder's adversarial term carries any part of STREAM's margin. We substitute DINO ViT-S/8 rather than a PatchGAN because DINO ViT-S/8 is the discriminator RAE uses, and RAE is the pipeline STREAM's decoder is built on, which makes it the field-standard control for this architecture and a stronger baseline than PatchGAN from the same natural-image domain.

Table 22: Discriminator ablation on BRCA at 45k decoder steps, everything else identical.

Discriminator	rFID ↓	gFID ↓
Kaiko ViT-S/8 (deployed)	**3.52**	**6.86**
DINO ViT-S/8 (RAE's)	4.33	7.21

The domain-matched discriminator improves STREAM by 0.81 rFID and 0.35 gFID (Table 22), far less than STREAM's margin over the Euclidean baselines in Table 3, where RAE reaches 15.84 gFID against STREAM's 6.16, so the adversarial prior does not explain that margin. Giving RAE the Kaiko discriminator in place of its own DINO ViT-S/8, on identical UNI embeddings and BRCA, raises its gFID from 15.84 to 16.81, so the domain-matched discriminator gives RAE no advantage.

E.6 Reproducibility across sampling seeds

We ask how far the reported numbers move when the sampling seed is the only thing that changes.

We report SDs over sampling seeds. Each seed redraws $\mathbf{Z}_0 \sim \mathrm{Uniform}((\mathcal{S}^{d-1})^N)$ and re-runs generation and evaluation, while the trained weights are held fixed. We report gFID only, because rFID reconstructs a fixed set of real images and is seed-independent.

Five independent sampling seeds at 10k samples each estimate the spread, and because FID depends on sample count, the reported values are the 50k-sample values with the seed SD beside them. Table 23 reports the four ablation cells of Table 4 under this protocol. We additionally re-measure

end-to-end at the full 50k-sample count over three seeds, which is the stricter check and is reported in Table 24.

Table 23: gFID at 50k samples (as in Table 4) ± SD over 5 sampling seeds of 10k samples each, for the four cells of Table 4 at the 45k decoder budget on BRCA.

Ablation cell	gFID (50k) ± seed SD
No bridge, isotropic	9.07 ± 0.24
No bridge, anisotropic	8.27 ± 0.21
Bridge, isotropic	9.37 ± 0.15
Bridge, anisotropic (ours)	$\mathbf{6.86 \pm 0.18}$

Table 24: gFID re-measured end-to-end at the full 50k-sample count, mean ± SD over 3 sampling seeds, BRCA. All three models use the shared UNI encoder of Table 3. The RAE and SVG values differ from Table 3 because this is a three-seed re-measurement rather than the single run reported there.

Model	gFID at 50k, mean ± SD
STREAM	$\mathbf{6.12 \pm 0.04}$
RAE	15.88 ± 0.04
SVG	17.52 ± 0.14

RAE and SVG reproduce their reported means to within 0.15 gFID (Table 24). STREAM's three-seed 50k-sample mean of 6.12 ± 0.04 lies within 0.05 of the Table 3 value of 6.16. Its five-seed 10k-sample SD is 0.20. STREAM's seed statistics are measured on the deployed decoder at the full 90k decoder schedule. The four ablation cells of Table 23 are the only ones repeated across sampling seeds, and every other ablation cell is a single run.

F Downstream Evaluation of Generated Images

Distribution metrics cannot on their own establish diagnostic utility. Since STREAM is unconditional and there are not many public labelers (i.e. classification models) that are robust and clinically validated enough for a downstream evaluation, the evaluations are therefore built either on the generated pixels directly or on a label-free pretraining setting. Appendix F.1 and Appendix F.2 both corroborate that STREAM-generated images are indeed biologically faithful.

F.1 Evaluation: Nuclear morphometry

We run HoVer-NeXt [3], a published pretrained nuclei segmentation-and-classification model, over two disjoint sets of 50k real COADREAD patches (real A and real B) and 50k STREAM-generated patches (∼2.6–2.8M nuclei per arm). Shape descriptors follow scikit-image `regionprops` conventions [60], with perimeter by the Crofton estimator; circularity follows the definition of Janssen et al. [26]. Nuclei per tile is HoVer-NeXt's instance count.

Real A and real B are drawn without overlap from the same pool of 1M real tiles, so real A against real B gives a tile-disjoint agreement floor.

Table 25: Nuclear morphometry on COADREAD, mean values over ∼2.6–2.8M nuclei per arm.

Mean, COADREAD	real A	real B	STREAM
nuclei per tile	51.87	51.49	55.36
area (px^2)	195.4	197.1	195.2
perimeter (px)	49.61	49.78	49.91
circularity	0.840	0.840	0.846
eccentricity	0.741	0.742	0.737
solidity	0.935	0.935	0.937
major axis (px)	19.11	19.18	19.13
minor axis (px)	11.14	11.17	11.36

Table 26: Downstream linear-probe evaluation of frozen DINO ViT-S/16 encoders pretrained only on 1M STREAM-generated tiles vs. 1M real TCGA tiles per organ. AUROC on out-of-distribution SPIDER/MHIST binary tasks. HP is hyperplastic polyp, SSA is sessile serrated adenoma, HG is high grade and LG is low grade. Rand. is a random-init control averaged over 3 network draws (5 for MHIST).

Linear-probe AUROC ↑	Rand.	**STREAM**	Real
SPIDER COADREAD Adenoma vs. Adenoca.	0.890	0.952	0.958
SPIDER COADREAD Adenoca. HG vs. LG	0.883	0.964	0.959
SPIDER COADREAD Adenoma HG vs. LG	0.668	0.951	0.943
SPIDER COADREAD HP vs. sessile serrated lesion	0.926	0.977	0.979
MHIST COADREAD HP vs. SSA	0.612	0.879	0.867
SPIDER BRCA DCIS HG vs. LG	0.976	0.998	0.998
SPIDER BRCA NST vs. Lobular	0.766	0.953	0.966
SPIDER BRCA In-situ vs. Invasive	0.935	0.994	0.994
SPIDER BRCA Benign vs. Malignant	0.737	0.979	0.983

Table 27: HEST spatial-transcriptomics regression on the same frozen encoders. Values are the average PCC over the top-50 highly variable genes. Rand. is a random-init control averaged over 3 network draws.

	Rand.		**STREAM**		Real	
HEST PCC ↑	COADREAD	BRCA	COADREAD	BRCA	COADREAD	BRCA
Ridge	0.144	0.163	0.284	0.313	0.280	0.333
XGBoost	0.128	0.145	0.247	0.307	0.269	0.283

As Table 25 shows, STREAM's mean sits within 2.0% of real on all seven per-nucleus shape descriptors and within 7.5% on nuclei per tile. Nuclear size and shape are the features a pathologist reads, so agreement at this level indicates that STREAM's samples are biologically faithful, while a blinded reader study with board-certified pathologists would be the ideal evaluation of actual clinical relevance.

F.2 Evaluation: Self-supervised pretraining and downstream evaluations

We believe that STREAM is best suited to this image augmentation setting that needs no labels and can help train foundation models by increasing the pretraining corpus with STREAM-generated synthetic data. We mock this by training a proof-of-concept DINO ViT-S/16 [9] model from scratch for 75 epochs under identical hyperparameters on 1M STREAM-generated tiles per organ (BRCA and COADREAD). As a fair comparison, we also train the same model separately on 1M real tiles. Then, we evaluate the two on downstream tasks with a random-init ViT-S/16 as the control. The two DINO models act as encoders and they are frozen and probed identically so only embedding quality is compared. We report linear-probe AUROC on SPIDER [46] with no slide contamination between train and test splits, plus MHIST [65] for COADREAD. Both probe sets are out-of-distribution (no TCGA overlap), while both encoders, real or STREAM-generated, see only TCGA tiles during DINO training. The nine linear probes are reported in Table 26. It is safe to conclude that an encoder pretrained only on STREAM-generated images is essentially on par with one pretrained on real images.

Classification is close to saturated in histopathology, as the high AUROC values in Table 26 show, so we run a more difficult downstream task on the same pretrained encoders. HEST [27] pairs hematoxylin and eosin (H&E) tiles with matched spatial transcriptomics; the task is to predict, from a tile's frozen embedding alone, the expression of each of the top-50 highly variable genes. We fit a ridge regression and a gradient-boosted regressor and report the average Pearson correlation coefficient (PCC), per gene and averaged over the 50 genes and the folds, on the official uncontaminated split, with results in Table 27. Across both datasets (BRCA and COADREAD) and downstream model (Ridge, XGBoost), STREAM again performs on par with real on average. We believe that this is stronger evidence of the biological fidelity of STREAM-generated synthetic images, because morphology only predicts correct expression when the embedding has captured underlying biologically meaningful features.

While these two downstream evaluations are not clinically validated, STREAM offers the first steps of an actual deployment scenario by addressing two constraints that van der Laak et al. [59] document in computational pathology. Privacy and approval requirements narrow the reuse of existing data for the development of artificial-intelligence models and raise the cost of assembling enough cases, and collecting annotations remains an obstruction to scaling algorithms. A generative model like STREAM transfers morphology without transferring any patient's tissue, which addresses the first. Because annotation is the scarce resource, an unconditional generative model matches the second constraint rather than falling short of it, where a conditional one would require exactly the labels that are hardest to obtain. The self-supervised result above is that workflow end to end, with a receiving institution pretraining an encoder on transferred synthetic tiles and reaching real-data quality while holding neither a label nor a patient slide.

G Ablations and Experiments on Decoder Noise and the Spectral Basis

G.1 Ablation on Low-Energy Noise Magnitude σ_L

We test how much noise the low-energy directions need by sweeping σ_L on BRCA at 45k decoder training steps, holding everything else at the deployed configuration. The deployed 0.02 gives the best gFID at 6.86, while 0.01 reaches a better rFID at 2.41 and 0.03 is worse on both (Table 28). Raising σ_L trains the decoder to denoise larger perturbations, so its reconstruction of clean inputs degrades and rFID rises throughout the sweep. gFID does not follow, and the deployed value is at the optimal minimum point.

Table 28: Decoder σ_L ablation on BRCA at 45k decoder steps. Bridge ($\sigma_{\max}$=0.01) and σ_H=0.002 held fixed.

σ_L	rFID ↓	gFID ↓
0.01	**2.41**	7.39
0.02 *(deployed)*	3.52	**6.86**
0.03	5.12	7.97

G.2 Ablation on Energy Threshold τ

We test where to cut the spectrum by ablating the energy threshold τ, which sets k^* and therefore the split between U_H and U_L, at the matched 45k decoder budget with $\sigma_H = 0.002$, $\sigma_L = 0.02$ and the deployed bridge held fixed. Lowering τ from 0.90 to 0.80 costs 2.29 rFID and 1.94 gFID (Table 29). The likely mechanism is that a lower τ shrinks k^* and moves the least energetic of the retained directions into the low-energy subspace, where they meet the far larger σ_L.

Table 29: Energy-threshold ablation on BRCA at 45k decoder steps.

τ	rFID ↓	gFID ↓
0.90 *(deployed)*	**3.52**	**6.86**
0.80	5.81	8.80

G.3 Ablation on Left vs. Right Singular Vectors

We test whether the decoder's noise should be shaped by the left or the right singular vectors of the Jacobian. Because the noise enters the decoder's input space, the right singular vectors look like the natural choice. J is an endomorphism of one token's tangent space, so both bases span that space and both select $k^* = 236$ at $\tau = 0.90$, differing only in which 236 directions they pick. Forward-mode JVPs apply J rather than $J^\top$, so the deployed cache holds the left singular vectors, and $\mathrm{range}(J)$ is where the velocity field's local variation lives, which makes U_H the directions of largest field response.

Retraining a full anisotropic decoder on the right singular vectors while holding everything else fixed costs 1.15 rFID and 1.25 gFID (Table 30). On the right basis too, shaping the noise by the spectrum beats isotropic noise, by 1.84 rFID and 1.26 gFID, and both spectral partitions inject less

total variance than the isotropic baseline rather than more. Spectral shaping therefore gives most of the rFID gain, and the shaping and the switch to the left basis give similar gFID gains.

Table 30: Left vs. right singular vector ablation on BRCA at 45k decoder steps.

Decoder noise basis	rFID ↓	gFID ↓
Left singular vectors, forward-mode (deployed)	**3.52**	**6.86**
Right singular vectors, reverse-mode	4.67	8.11
Isotropic, 0.02 on every direction	6.51	9.37

G.4 Projected Energy of the Prediction Residual

The anisotropic allocation comes from the DiT's learned velocity field. The SVD of the field's Jacobian separates the directions of largest response, U_H, from the directions of smallest response, U_L (Section 4.3). The decoder trains under only $\sigma_H = 0.002$ along U_H, while U_L absorbs the large robustness budget $\sigma_L = 0.02$. We measure where the generator's error lands, which is a separate question. The allocation does not require that error to concentrate in U_L.

For each of 2,000 BRCA tiles we encode z_1 with UNI, form the deployed bridge interpolant x_t from x_0 drawn uniformly on $(\mathcal{S}^{d-1})^N$, and query the deployed DiT once. The residual $e = \Pi_{z_1}(\hat{z}_1 - z_1)$ is the tangent part of this teacher-forced prediction error. We parallel-transport e into the tangent space of the tile's assigned centroid, where the deployed t=0.99 cache is orthonormal. For token n, $E_{H,n}$ is the energy of e_n along the top-k cached left singular vectors of that (centroid, token), and $E_{L,n} = \|e_n\|^2 - E_{H,n}$ is the energy in their orthogonal complement. E_H and E_L sum these over the 256 tokens and average over tiles. The per-dimension density ratio

$$\rho = \frac{E_L / \dim U_L}{E_H / \dim U_H}$$

equals 1 when the residual spreads evenly across directions. The tiles come from the training distribution, since no held-out split exists for the deployed BRCA DiT.

Table 31: Projected energy of the DiT's teacher-forced endpoint residual on BRCA, n=2,000 tiles, UNI encoder, projected onto the deployed t=0.99 cache. E_H and E_L are per-tile sums over the 256 tokens averaged over tiles; the U_L share is $E_L/(E_H{+}E_L)$ and the even-spread share is $\dim U_L/1023$; ρ is the per-dimension density ratio, equal to 1 under an even spread. The upper block varies the partition at the deployed anchor and the lower block varies the anchor at the deployed partition. The per-token row uses each (centroid, token)'s own k^*, and its dimensions are means over the 2,000 tiles.

	$\dim U_H$	$\dim U_L$	E_H	E_L	share in U_L	even spread	ρ
Partition varied at the deployed anchor t=0.99							
k=32	32	991	0.0563	0.6686	92.2%	96.9%	0.383
k=64	64	959	0.0931	0.6318	87.2%	93.7%	0.453
k=128	128	895	0.1474	0.5775	79.7%	87.5%	0.560
k=236 (deployed)	236	787	0.2254	0.4995	68.9%	76.9%	0.664
k=380	380	643	0.3218	0.4031	55.6%	62.9%	0.740
per-token k^*	237.9	785.1	0.2260	0.4989	68.8%	76.7%	0.669
Anchor varied at the deployed partition k=236							
t=0.5	236	787	4.1103	12.4434	75.2%	76.9%	0.908
t=0.7	236	787	2.4151	7.7522	76.2%	76.9%	0.963
t=0.9	236	787	0.9256	3.2628	77.9%	76.9%	1.057

The residual does not concentrate in U_L (Table 31). At the deployed anchor ρ stays below 1 at every partition, the deployed one included, so the residual is denser per dimension in U_H. At the deployed partition ρ sits near 1 away from that anchor. A density below 1 does not undercut the allocation. To first order the decoder responds to the total energy in each subspace rather than to its density, and the larger total lands in U_L, where the large noise goes.

G.5 Directional-Lipschitz Probe (Extended)

Table 32 runs the directional-Lipschitz probe of Table 5 on the Euclidean RAE pipeline, which removes the spherical geometry while keeping the latent dimension, the decoder family and the noise magnitudes. On the sphere, the no-bridge k=64 row of Table 5 reverses the same way as the bridge row, from 0.62 isotropic to 2.50 anisotropic, so the reversal belongs to the spherical geometry and the anisotropic allocation, and the bridge sharpens it rather than creating it. The RAE ratios show no directional split, and anisotropic training does not create one. The same absence explains the null result of Appendix E.3, because without the spherical geometry there is no U_H/U_L split for the anisotropic decoder to act on.

Table 32: Directional-Lipschitz ratio $R_{\text{dec}} = L^{\text{dec}}_{U_H}/L^{\text{dec}}_{U_L}$ for the Euclidean RAE pipeline on BRCA, n=2,000 tiles, as the control for Table 5. Bridge rows use the stochastic bridge and no-bridge rows do not; k^*=380 is RAE's own τ=0.90 energy cutoff, the counterpart of STREAM's 236. These rows use a plain LightningDiT-XL without the DDT head so that latent dimensions match STREAM and the SVD extraction is clean.

Pipeline	Isotropic	Anisotropic
RAE, no bridge (k=64)	0.975	0.982
RAE, no bridge (k^*=380)	0.992	0.996
RAE, bridge (k=64)	0.9797	0.9863
RAE, bridge (k^*=380)	0.9928	0.9979

G.6 Directional Response in Pixel Space

Figures 6 and 7 show in pixels the ablation of Table 4 and the directional-LPIPS values of Table 5; Fig. 6 is the full version of Fig. 3, with the 1× and 8× columns and the $|\Delta|$ rows added. We take 16 latents generated by the bridge DiT of Table 4, and the figures display one of them. For each latent we perturb all 256 tokens inside a subspace of that latent's own per-(centroid, token) frame and renormalize each token onto $\mathcal{S}^{d-1}$. The perturbation is drawn once, and the isotropic and anisotropic bridge decoders of Table 4, both at 45k steps, decode the same perturbed point. Figure 6 perturbs the leading 64 directions of the frame, the k=64 partition of Table 5 rather than the deployed k^*=236 one. Figure 7 perturbs U_L, the 787-dimensional orthogonal complement of the deployed U_H in the 1023-dimensional tangent space. We draw it as Table 5 does, projecting an ambient Gaussian onto the tangent space at each token and removing its U_H component.

The amplitude unit is $\epsilon_H = \sigma_H\sqrt{k^*} = 0.0307$, the per-token displacement that the deployed covariance of Eq. (6) places inside U_H. Column m× displaces every token by $m\,\epsilon_H$ in its tangent space. The deployed displacement inside U_L, $\sigma_L\sqrt{787} = 0.56$, falls between the 8× and 32× columns. The statistic is the mean absolute pixel difference from each decoder's own clean decode on the 0–255 scale. It is not LPIPS, and its magnitude is not comparable with R_{dec}. For display, each $|\Delta|$ panel is divided by its own displacement $m\,\epsilon_H$ and multiplied by one factor shared by both figures, so black is no change and brightness compares across the two figures.

On the top-64 block the anisotropic decoder clearly separates from the isotropic one and responds more strongly to the perturbation (Fig. 6). On U_L, the control, both decoders train under the same σ_L and behave the same, as expected (Fig. 7). We read the separation as an empirical directional-robustness effect, not as an established principle. Nothing is visible at the deployed amplitude (1×); in the H&E rows the separation becomes visible from 32× onward, and the amplified $|\Delta|$ rows show it at every amplitude. The figures cover one cell and 16 generated latents, so they are descriptive only.

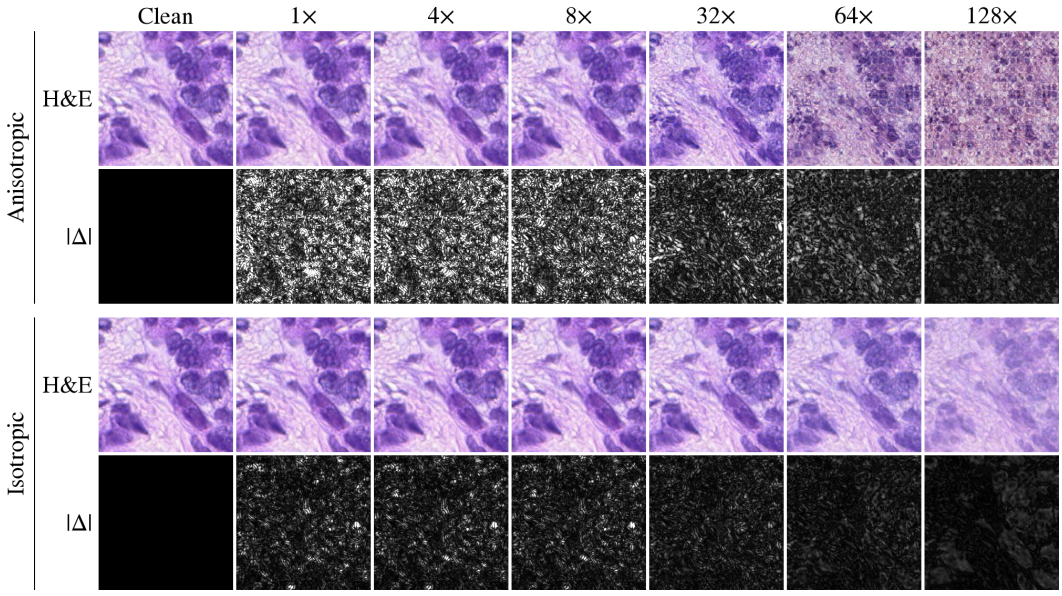


Figure 6: Pixel response of the anisotropic (upper block) and isotropic (lower block) decoders to a perturbation inside the top-64 subspace, BRCA. In each block the H&E row holds the decodes and the $|\Delta|$ row their absolute difference from the clean decode divided by the displacement $m\,\epsilon_H$ (response per unit displacement; black is no change), for the clean latent and six amplitudes $m\times$.

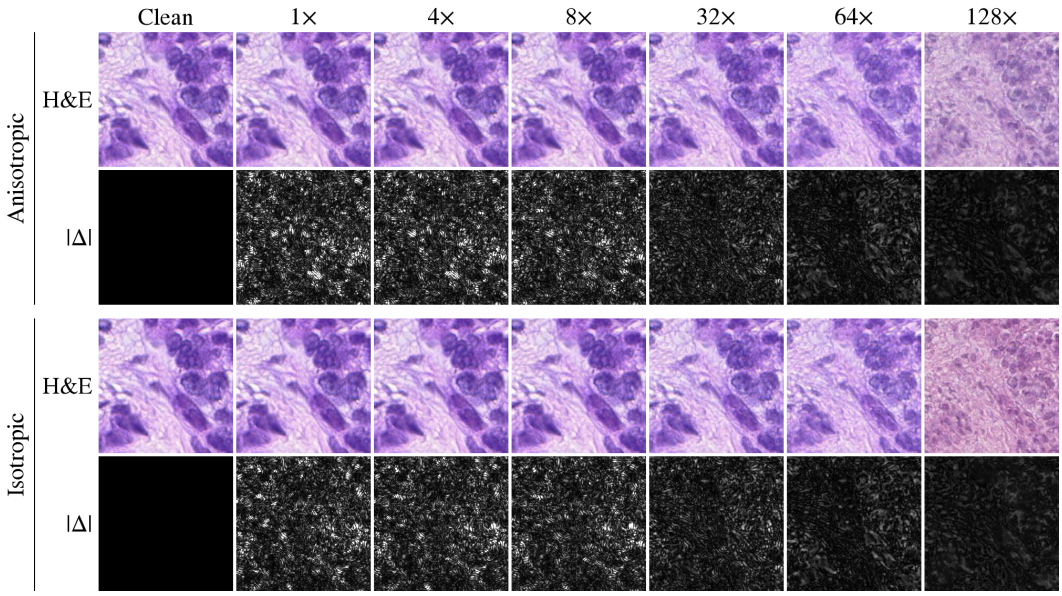


Figure 7: The same measurement as Fig. 6 with the perturbation inside U_L, the orthogonal complement of the deployed U_H.

H Encoder Quality Dependence

This appendix tests whether the DiT or the decoder fails under the domain-mismatched encoder of Table 6, DINOv2-L on BRCA, and places the failure in the decoder rather than the DiT. It also measures how the spectral basis depends on its time anchor and reruns STREAM on a second in-domain encoder.

H.1 DiT vs. Decoder Isolation via Latent-Space Fréchet Distance (FD)

We test whether the DINOv2-L DiT itself fails to generate plausible latents, which would place the failure of Table 6 in the generator rather than the decoder. For each encoder we compute the FD between Gaussian fits to 1,000 DiT-generated latents and to the corresponding real test latents, on mean-pooled tokens in that encoder's own feature space. It is the FID statistic applied to encoder features. Generation uses the deployed integrator at 25 steps, so no decoder enters the measurement.

The UNI DiT reaches an FD of 0.030 and the DINOv2-L DiT 0.032, so the DINOv2-L DiT generates distributionally correct latents at the level of mean-pooled Gaussian statistics; the test does not probe per-token structure. The missing gain of Table 6 also appears in rFID, which decodes real latents and involves no DiT: anisotropic noise moves DINOv2-L's rFID from 7.67 to 7.71, against 6.51 to 3.52 for UNI (Table 4). The end-to-end failure of Table 6 therefore lies in the decoder rather than the DiT. We attribute it to the out-of-domain encoder: the DiT learns DINOv2-L's latent distribution as faithfully as UNI's, but these latents encode natural-image features, which the decoder maps less well to high-fidelity histopathology images.

H.2 Time-Anchor Sensitivity of the Spectral Basis

We measure how far the spectral basis moves when its time anchor changes. The deployed cache is extracted at the single anchor $t = 0.99$ (Appendix D) and is neither time-averaged nor recomputed during sampling. The dominant subspace is not invariant in t, and we do not claim that it is.

We re-extract the cache at the adjacent anchor $t = 0.9$ on the same 50 centroids, with only the anchor differing. For each (centroid, token) we compare the top-k=64 subspaces of the two anchors by their principal-angle alignment [7] $A = \|Q_1^\top Q_2\|_F^2/k \in [0, 1]$, where Q_1, Q_2 are orthonormal bases of the two subspaces and $A = 1$ for identical subspaces, and average A over tokens. The cross-centroid reference is the same token-matched statistic between two different centroids at the same anchor t=0.99, averaged over all $\binom{50}{2} = 1{,}225$ pairs; it is the variation the K=50 cache is built to absorb. The chance level is $k/(d-1) = 0.063$, the expected alignment of two independent, uniformly random k-dimensional subspaces of the $(d-1)$-dimensional tangent space.

Table 33 reports three alignments. The first row changes only the anchor, from t=0.99 to t=0.9, at a fixed centroid and token; the second changes only the centroid, at the fixed anchor t=0.99 and a fixed token; the third is the chance level. The adjacent-anchor alignment is $A = 0.297$ $[0.283, 0.312]$, well above the chance level of 0.063, so the two anchors share part of their dominant subspace. The whole interval, however, lies below the cross-centroid reference of 0.529. The anchor thus matters more than the cross-centroid variation the cache is built to absorb.

Table 33: Top-k=64 subspace alignment A of the UNI cache on BRCA. Intervals are 95% percentile bootstrap over the 50 centroids.

Comparison	A
Adjacent anchors, t=0.9 vs. t=0.99, same centroid	0.297 [0.283, 0.312]
Different centroids, same anchor t=0.99 (cross-centroid reference)	0.529
Independent uniformly random subspaces (chance level, $k/(d-1)$)	0.063

Limitations. The single-anchor cache is therefore an approximation and a limitation. We have not measured stability along the sampler's own trajectory, which is a different quantity. The extraction linearizes at clean centroids, and measuring it would require re-linearizing at the sampler's iterates, where the base point moves as well as the time label. However, the single-anchor cache still buys a measured gain. At the matched 45k decoder budget the deployed basis improves gFID over isotropic noise (Table 4), and swapping it for a different but equally valid basis costs 1.25 gFID (Appendix G.3).

H.3 A Second In-Domain Encoder: Virchow2

Table 6 replaces UNI with a domain-mismatched encoder. To test whether STREAM's gain is specific to UNI's spectrum rather than to in-domain histopathology pretraining generally, we rerun

the full pipeline with Virchow2 features and hold the dataset, the tiling and the 45k decoder budget fixed.

Table 34: Second in-domain histopathology encoder, BRCA at the matched 45k decoder budget. Only the within-encoder isotropic-to-anisotropic contrast is a valid comparison; see the confounds below.

Configuration	rFID ↓	gFID ↓
Virchow2, bridge + anisotropic (STREAM)	**6.37**	**10.48**
Virchow2, bridge + isotropic	11.28	14.75
UNI, bridge + anisotropic (STREAM)	**3.52**	**6.86**
UNI, bridge + isotropic	6.51	9.37

STREAM's bridge and anisotropic decoder transfer to Virchow2 (Table 34). With the bridge in both arms, the anisotropic decoder improves on its isotropic control by 4.91 rFID and 4.27 gFID. In relative terms the gains are close to UNI's, 44% and 29% against 46% and 27%, so they are not specific to UNI's particular spectrum. They remain encoder-dependent, as DINOv2-L shows in Table 6, and Appendix H.4 states their scope.

Two confounds, both forced by the encoder. First, Virchow2 emits 1280-dimensional patch tokens, above the hidden size of 1152 of the deployed LightningDiT-XL (Table 13). RAE [77] reports that a DiT whose hidden size is below the token dimension hits a bottleneck, and Zheng et al. [77, Thm. 1] establishes a non-zero loss lower bound for diffusion models of the form $Bf(Ax_t, t)$ when the hidden dimension satisfies $d < n$. This run therefore uses LightningDiT-1.0B (1.03B parameters against 676M) with a correspondingly larger SVD cache and decoder. Only the within-encoder contrast is valid, and the absolute Virchow2 and UNI rows are not comparable. UNI's d=1024 sits below 1152, so the deployed configuration is unaffected. Second, Virchow2's patch size is 14 rather than 16, so each tile is resized from 256 to 224 pixels to keep 256 tokens per tile. The training and evaluation data are otherwise identical, but Virchow2 sees each tile at lower resolution. We do not re-tune STREAM's hyperparameters for Virchow2 and expect their optimal values to differ.

H.4 Scope of the Encoder Dependence

The anisotropic decoder's gain transfers between two in-domain histopathology VFMs, UNI and Virchow2, at a similar relative size (Table 34), but not to the natural-image DINOv2-L (Table 6). We do not claim that the anisotropic decoder is effective independently of the encoder. With DINOv2-L the DiT still generates distributionally correct latents (Appendix H.1), while anisotropic noise leaves reconstruction unchanged, so the encoder dependence enters at the decoder rather than the generator. We have no statistic that predicts in advance whether a new encoder or modality benefits; the isotropic-versus-anisotropic decoder contrast at the matched 45k budget is the test. We regard domain matching as the most reliable way to obtain the gain. If a target domain has no in-domain encoder of this quality, we expect the anisotropic decoder's benefit to degrade.

I Image Gallery

I.1 Reconstruction comparison on BRCA

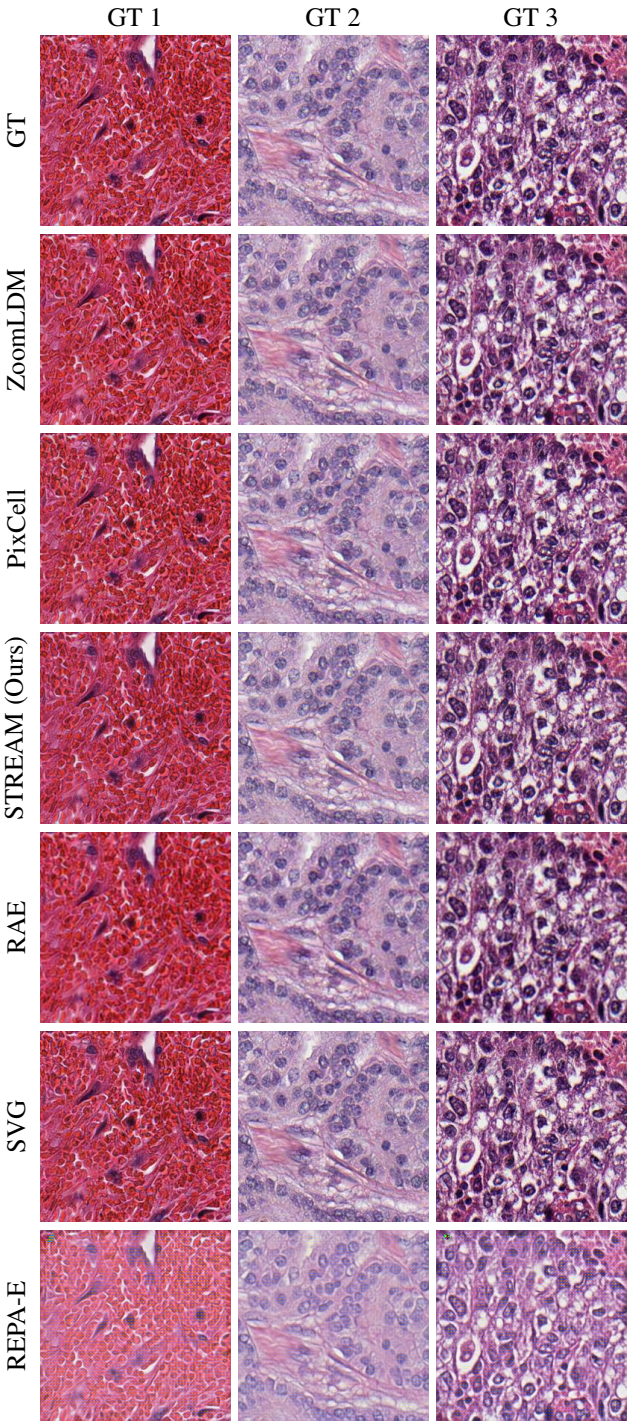


Figure 8: Reconstructions on BRCA from ZoomLDM, PixCell, STREAM (Ours), RAE, SVG and REPA-E, taken from the rFID evaluation, with GT tiles in the top row. The tiles are selected by per-image LPIPS ranking of the GT–reconstruction pairs to highlight perceptual differences across models. The six models shown all use the shared UNI encoder of Table 3.

I.2 Reconstruction comparison on COADREAD

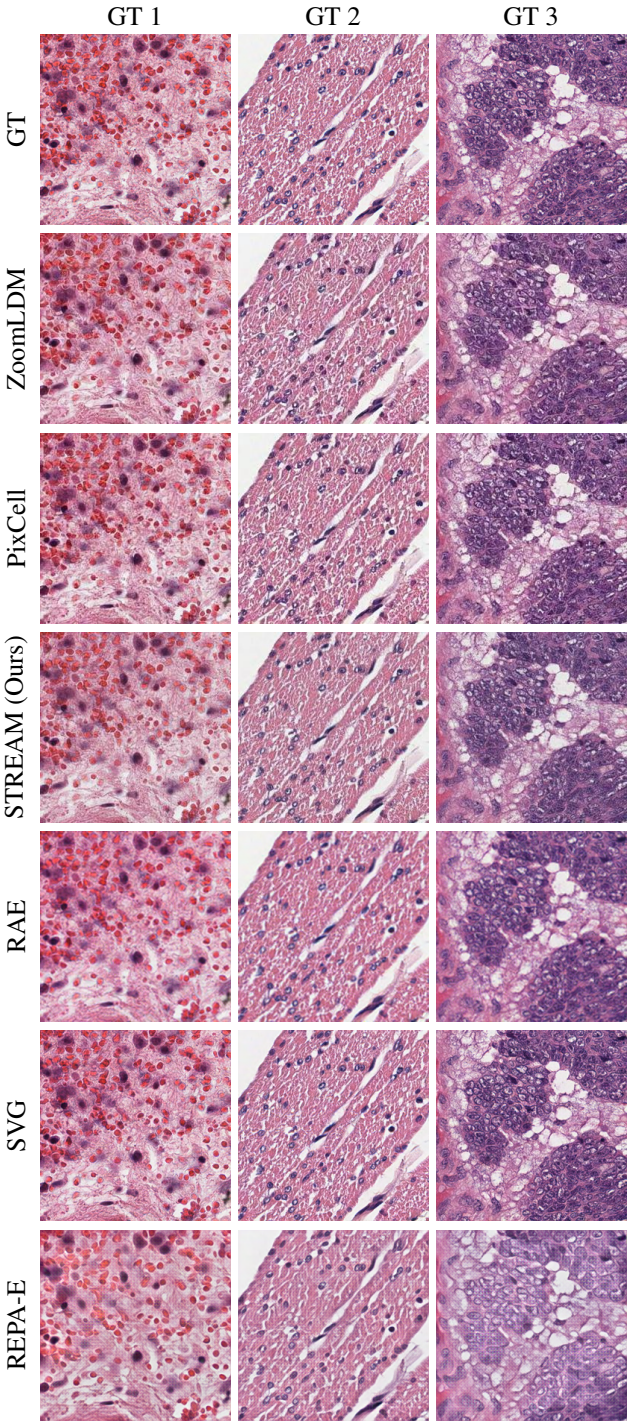


Figure 9: Reconstructions on COADREAD from ZoomLDM, PixCell, STREAM (Ours), RAE, SVG and REPA-E, taken from the rFID evaluation, with GT tiles in the top row. The tiles are selected by per-image LPIPS ranking of the GT–reconstruction pairs to highlight perceptual differences across models. The six models shown all use the shared UNI encoder of Table 3.

I.3 Generation samples on BRCA

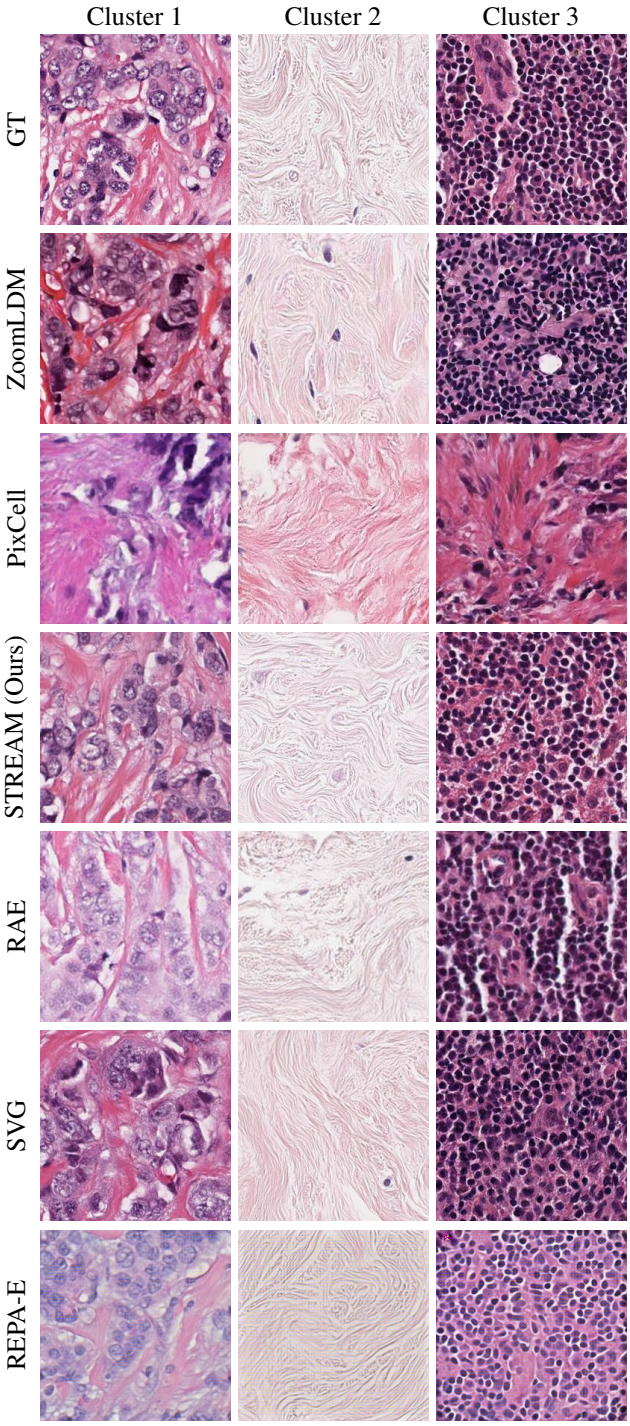


Figure 10: Generation samples on BRCA. Unconditional samples used for gFID evaluation from ZoomLDM, PixCell, STREAM (Ours), RAE, SVG and REPA-E, with GT tiles in the top row. Each column corresponds to a representative tissue cluster obtained by K-means on Virchow2 features of the GT set. For each cluster and model we display one of the three samples whose Virchow2 feature lies closest to the cluster centroid. The six models shown all use the shared UNI encoder of Table 3.

I.4 Generation samples on COADREAD

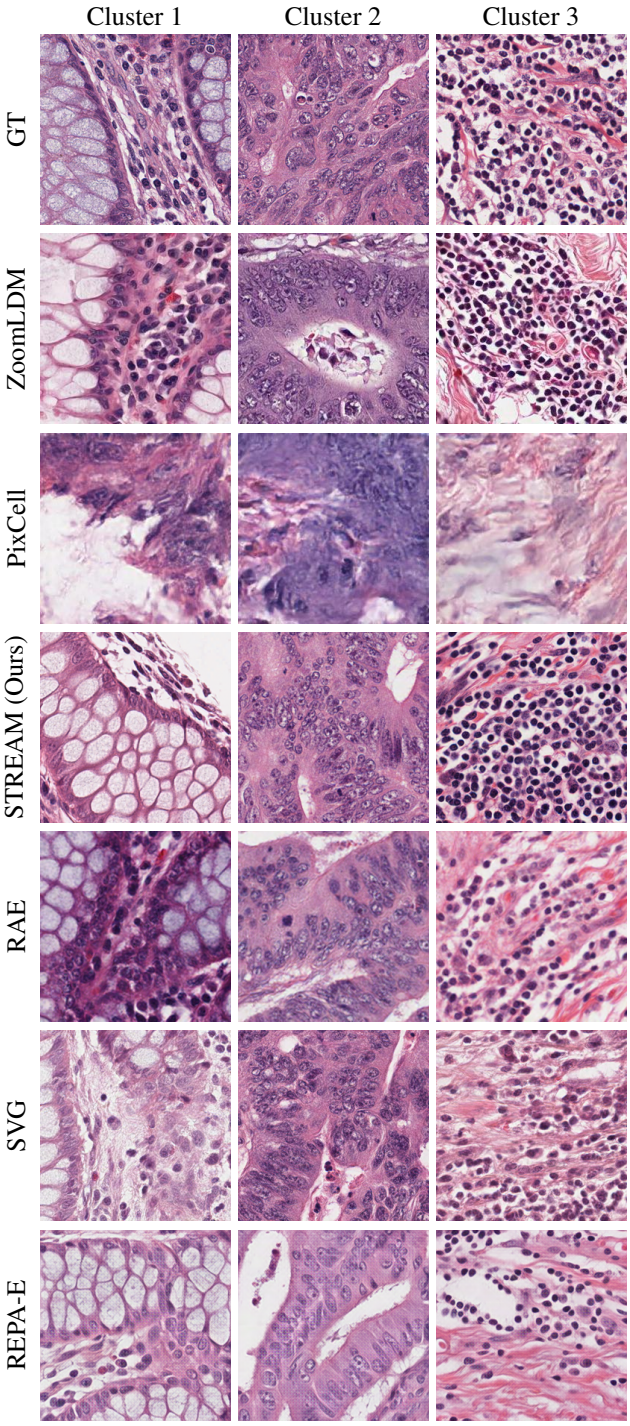


Figure 11: Generation samples on COADREAD. Unconditional samples used for gFID evaluation from ZoomLDM, PixCell, STREAM (Ours), RAE, SVG and REPA-E, with GT tiles in the top row. Each column corresponds to a representative tissue cluster obtained by K-means on Virchow2 features of the GT set. For each cluster and model we display one of the three samples whose Virchow2 feature lies closest to the cluster centroid. The six models shown all use the shared UNI encoder of Table 3.

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 state three contributions: (1) empirical motivation for Riemannian flow matching in histopathology via the conditioning-dominated-diversity analysis (Section 3), (2) bridge-type stochastic perturbation with per-token rectifiability under the hypotheses of Theorem 1 ($\sigma_{\max} > 0$ and a coupling with no mass on the antipodal locus), and (3) spectral-informed anisotropic decoder *training* via the left-singular basis of the velocity-field Jacobian (supported empirically by the ablation in Table 4). The abstract's performance claim concerns *generation* quality, namely gFID and the histopathology-specific metrics, and Table 3 supports it with every method on the same UNI encoder. Reconstruction is not claimed, since SVG leads rFID on all three datasets (Tables 3 and 18). The downstream claim, that a self-supervised encoder pretrained only on STREAM samples performs on par with one pretrained on real tiles, is supported by Section 5.2 and Appendix F. Section 7 states that Theorem 1 holds at the population level and that the account of the superadditive mechanism is mostly empirical. Proposition 3 states the cross-token hypothesis (CT), Appendix H.2 the single-anchor limitation, and Section 6 and Appendix H.4 the encoder dependence. Appendix F states that distribution metrics cannot on their own establish diagnostic utility.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 states that Theorem 1 holds at the population level and that our account of the superadditive mechanism is mostly empirical (Appendix G.4). Proposition 3 and its proof state hypothesis (CT) in full, note that it is assumed and not proved, and state that the $O(\sigma_{\max}^2)$ transport rate makes no finite-parameter guarantee at the deployed values. Appendix H.2 names the single time anchor t=0.99 of the spectral cache as an approximation and a limitation. Section 6 and Appendix H.4 scope the anisotropic decoder to in-domain encoders, and Appendix H.4 states that no encoder-independent effectiveness is claimed. Appendix F states that distribution metrics cannot on their own establish diagnostic utility and that the downstream evaluations are not clinically validated. Appendix F.1 names a blinded reader study with board-certified pathologists as the ideal evaluation of clinical relevance.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: All theorems, propositions and lemmas are numbered and cross-referenced. The main paper states Proposition 1 and Theorem 1, Appendix B.3 gives their complete proofs, and Appendix B states and proves the supporting lemmas and propositions. Assumptions are stated explicitly. Theorem 1 assumes $\sigma_{\max} > 0$ and a coupling giving no mass to the antipodal locus. The cross-token C^2 hypothesis (CT) is named in the statement of Proposition 3 and used nowhere else. Theorem 1 is proved directly on $\mathcal{S}^{d-1}$ and neither proves nor assumes a Riemannian extension of Hertrich et al. [21, Thm. 14] (Appendix B.2).

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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: The two-stage training pipeline is fully specified. Stage 1 trains the DiT with the stochastic bridge, and Stage 2 trains the decoder with anisotropic noise. Algorithms 1 and 2 detail Stage 1 and Stage 2 training, and Algorithm 3 details inference. Appendix D provides complete hyperparameters for both stages (Table 14), including learning rates, schedules, loss weights, SVD cache parameters (the time anchor t=0.99 and the deployed k^*=236) and adversarial training phases. Code and a public gallery of generated images are released (see the following item).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the model, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [Yes]

   Justification: The training, evaluation and spectral-cache code is publicly released under the CC BY-NC-SA 4.0 license at https://chokevin8.github.io/STREAM-Patho/, alongside a public gallery of STREAM-generated images. TCGA-BRCA and TCGA-COADREAD are publicly available via the Genomic Data Commons Data Portal, and SPIDER, MHIST and HEST are publicly available from their authors. UNI, UNI2-h and Virchow2 weights require an access request to their original authors.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: The two-stage pipeline, loss functions and training procedure are specified in Sections 4 and 5. Appendix D provides full hyperparameters in Table 14, covering the optimizer (AdamW), learning rates, schedules, batch sizes, loss weights, adversarial training phases, SVD cache settings and generation details (25 midpoint Euler steps, $\mathrm{NFE}{=}50$). Hyperparameters that are our own choices rather than inherited ones, namely the bridge scale $\sigma_{\max}$, the energy threshold τ, the low-energy noise σ_L, the noise-schedule shape and the discriminator, are individually ablated in Appendices C.1, E.5, G.1 and G.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: Appendix E.6 gives the seed protocol and the multi-seed statistics. The reported gFID values are the 50k-sample values, and their SDs come from five independent sampling seeds at 10k samples each. All four cells of the main ablation (Table 4) carry five-seed SDs in Table 23. rFID reconstructs a fixed set of real images and is seed-independent by construction. Training-seed variance is not measured, and every ablation cell outside Table 4 is a single run.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: STREAM is trained on one node of 8× NVIDIA H200 GPUs (Section 5.1 and Appendix D.1). Stage 1 (DiT) runs 95K steps at batch size 1024, and Stage 2 (decoder) runs 90K steps at batch size 512. Appendix D.1 gives the per-stage wall-clock and GPU-hour breakdown ($\sim$545 GPU-hours in total, of which the STREAM-specific spectral stage takes 69 GPU-hours, 12.7%). Appendix D.2 gives per-image inference FLOPs and latency at NFE=50 on a single NVIDIA H200.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 uses publicly available medical imaging data (TCGA-BRCA, TCGA-COADREAD, SPIDER, MHIST and HEST, all de-identified) and does not involve human subjects, deception, or harmful applications. Synthetic histopathology images are generated for research purposes (data augmentation, model training) and are not intended for clinical diagnosis. Appendix F states that distribution metrics cannot on their own establish diagnostic utility and that the downstream evaluations are not clinically validated.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Positive impacts include addressing data scarcity and privacy in computational pathology and enabling label-free cross-institutional pretraining of diagnostic models without patient data sharing (Section 5.2, Appendix F.2). On the negative side, the risk of misleading synthetic pathology images is mitigated by the specialized domain and the 256×256 patch resolution, which is insufficient for whole-slide-image-based clinical diagnosis. All released images are labeled as synthetic at the point of release.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: [Yes]

    Justification: The camera-ready release comprises code and a public gallery of STREAM-generated images. The residual misuse risk is low relative to general-purpose image generators. We nevertheless apply three safeguards. Every released image is explicitly labeled as machine-generated at the point of release. The code is released under the non-commercial CC BY-NC-SA 4.0 license. No slide identifier or protected health information is contained in the released artifacts, which derive from de-identified public TCGA and SPIDER data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: All existing assets are cited at first use. The encoders and discriminators are UNI and UNI2-h [13], Virchow2 [78], DINOv2 [47], DINOv3-L [54], Kaiko ViT-S/8 [29] and DINO ViT-S/8 and ViT-S/16 [9]. The pathology foundation models are used under their authors' licenses, which we observe. The datasets are TCGA-BRCA [58] and TCGA-COADREAD [57], which are publicly available under the National Institutes

of Health Genomic Data Sharing Policy, and SPIDER (breast, colorectal and skin subsets) [46], MHIST [65] and HEST [27], each used under the license its authors publish. The baseline methods and code are ZoomLDM [74], PixCell [73], RAE [77], SVG [52], REPA-E [35] and LightningDiT [71]. The analysis tools are HoVer-NeXt [3] (Tiny model, Lizard-Mitosis weights) and scikit-image [60], whose `regionprops` conventions we follow for the morphometry descriptors.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: We release the STREAM training, evaluation and spectral-cache code and a public gallery of STREAM-generated histopathology images at `https://chokevin8.github.io/STREAM-Patho/`. Documentation is provided alongside. It covers the training recipe and full hyperparameters (Appendix D, Table 14), the encoder and dataset each artifact is produced from, exact commands and environment specifications, and the license. Every released image is labeled as machine-generated. No new dataset of real tissue is released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: No crowdsourcing or human subjects research is involved. In particular, the paper includes no reader study with human pathologists and names a blinded reader study as the ideal evaluation of clinical relevance (Appendix F.1).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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 subjects research is involved. The TCGA, SPIDER, MHIST and HEST datasets are de-identified and publicly available.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: [N/A]

    Justification: LLMs are not used as a component of the core methodology. The research concerns Riemannian flow matching for histopathology image generation, not language models.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.