DENOISING STRAIN: A GEOMETRIC ORDER PARAMETER THAT REVEALS LEARNING AHEAD OF EVAL METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models are evaluated through denoising loss and sample-quality metrics such as Fréchet Inception Distance (FID). These scores track performance, but do not directly show the generative structure taking shape during learning. We introduce *Denoising Strain*, a geometric description of how each complete sampler step contracts, amplifies, and transports nearby perturbations. Separating overall scale from directional stretching gives a tensor order parameter: it vanishes when all directions are treated equally and develops structure as preferred directions form. The discovered axes have measurable consequences for generation. On pretrained animal-face and human-face models, leading-axis interventions produce median final-image separations about 28 and 24 times those of equal-norm random controls. A complementary study follows directions through denoising and finds stronger image-feature responses along the tracked axes than along random controls. We also track this geometry along uninterrupted MNIST and CIFAR-10 training runs, where initially near-equal stretches differentiate and expanding directions appear. On CIFAR-10, expansion and anisotropy are detected 250–500 times earlier in training exposure than the checkpoint completing half of the log-FID improvement from initialization to the best evaluated checkpoint. Denoising Strain thus provides an order parameter that links how diffusion models generate to how that ability develops during training.

1 INTRODUCTION

Training a diffusion model is usually monitored through denoising loss and sample-quality scores such as Fréchet Inception Distance (FID) (Ho et al., 2020; Karras et al., 2022; Heusel et al., 2017). These scores summarize changes in performance. To understand how that performance develops, we study how the sampling dynamics change during learning.

Our starting point is a simple geometric picture. Place a tiny ball of perturbations around an intermediate sampling state. To first order, the next step turns it into an ellipsoid (Figure 1). Shorter axes mean that differences shrink; longer axes mean that they grow. The axes also turn as denoising proceeds, carrying local differences toward different completed images. This changing geometry describes which alternatives the sampler suppresses, which it amplifies, and how it transports them.

We call this description *Denoising Strain*. We differentiate the complete numerical sampler step and measure its directional stretches on a logarithmic scale. Removing the common scale leaves an anisotropy tensor, which we propose as an *order parameter*: it vanishes when no direction is preferred and develops structure as directions become differentiated. The full strain also records whether each direction contracts or expands, and its axes can be followed through denoising.

The geometry matters for what the model generates. On pretrained Animal Faces-HQ v2 (AFHQv2) and Flickr-Faces-HQ (FFHQ) models, perturbing the direction with the greatest one-step stretch, the leading strain axis, produces final-image separations about 28 and 24 times those of equal-norm random controls. The images differ in animal category, pose, and facial appearance. A larger study follows the local directions of the continuous denoising flow. It tests 100 held-out trajectories per dataset at all 11 measured states and three perturbation magnitudes. Dose-response slopes measured with the frozen DINOv2 image encoder (Oquab et al., 2024) exceed matched controls at every state.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

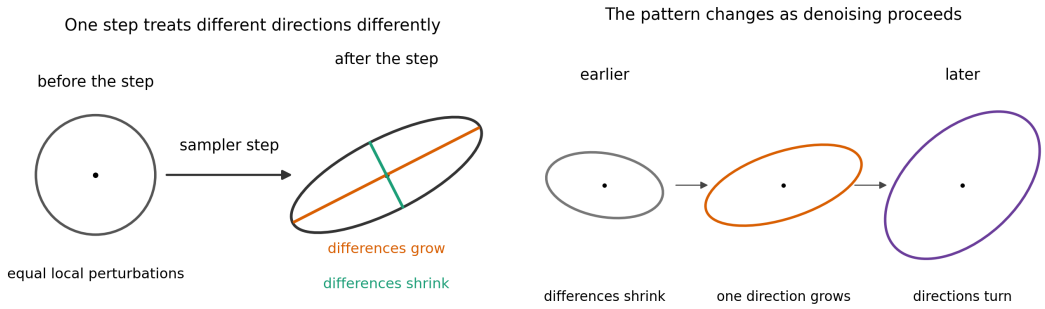


Figure 1: **Reading a sampler step through its effect on nearby states.** A small ball of perturbations becomes an ellipsoid. Its axis lengths show which differences shrink or grow; its orientation shows where those differences are carried. Denoising Strain separates the overall scale change from this directional structure.

We then follow this geometry through training. Across uninterrupted MNIST and CIFAR-10 training runs, the measured directions begin with nearly equal stretches and become differentiated, with some eventually expanding. On CIFAR-10, geometric events appear after 10,000–20,000 training images; the checkpoint completing half of the log-FID improvement from initialization to the best evaluated checkpoint arrives after 5 million images, 250–500 times later. On MNIST, different noise levels acquire their final-checkpoint geometry at different times. Later in training, feature-FID can worsen while geometric similarity, held-out loss, precision, and image morphology improve.

2 Measuring how a sampler step changes nearby states

To observe emerging structure, we need to separate directional organization from the sampler's overall contraction. The ellipsoid picture provides this separation: its axis lengths give directional stretches, their common scale gives overall contraction, and their differences reveal preferred directions. We derive this measurement for one sampler step and repeat it across training checkpoints.

2.1 Following a small perturbation

Let a trained diffusion model with parameters θ and a deterministic sampler define a sequence of transitions

$$x_{n+1} = \Phi_{\theta,n}(x_n), \qquad n = 0, \dots, N-1. \tag{1}$$

Here x_n is the intermediate image state at step n, N is the number of sampler transitions, and $d = \dim(x_n)$ is the state dimension. The map $\Phi_{\theta,n}$ includes the complete numerical update, including both predictor and corrector denoiser evaluations when using Heun's second-order numerical integration method. At the realized trajectory state x_n, its tangent map is

$$T_n = D_x \Phi_{\theta,n}(x_n). \tag{2}$$

Here D_x denotes the Jacobian with respect to the current state. For an infinitesimal perturbation δx_n, the next-step perturbation is $\delta x_{n+1} = T_n \delta x_n + o(\|\delta x_n\|)$. Thus T_n describes the local deformation produced by one complete numerical sampler step. We assume that T_n is locally nonsingular along the evaluated trajectories; otherwise zero singular values require a regularized or support-restricted logarithm.

2.2 Separating directional structure from overall contraction

We measure stretching by comparing perturbation lengths before and after the step. The right Cauchy–Green tensor $T_n^\top T_n$ removes the output-space rotation and retains the deformation seen from the current state. Its matrix logarithm defines the symmetric logarithmic strain

$$H_n = \frac{1}{2} \log(T_n^\top T_n). \tag{3}$$

Equivalently, if $T_n = U_n \Sigma_n V_n^\top$ is a singular value decomposition, then

$$H_n = V_n \log(\Sigma_n) V_n^\top, \qquad R_n = U_n V_n^\top, \tag{4}$$

where R_n is the orthogonal factor carrying the deformed frame into the next state's tangent space. The columns $v_{n,i}$ of V_n are unit input axes, and $H_n v_{n,i} = h_{n,i} v_{n,i}$ defines their log stretches. An input perturbation along $v_{n,i}$ is stretched by $\exp(h_{n,i})$ during the step. Negative $h_{n,i}$ means contraction, positive $h_{n,i}$ means expansion, and $h_{n,i} = 0$ is the local expansion boundary.

A sampler can contract every direction while still treating some directions very differently from others. We isolate that preference by subtracting the mean log stretch from H_n:

$$m_n = \frac{1}{d} \operatorname{tr}(H_n) = \frac{1}{d} \log |\det T_n|, \qquad Q_n = H_n - m_n I. \tag{5}$$

Here I is the identity matrix. The complete step can therefore be written as

$$T_n = R_n \exp(H_n) = R_n \exp(m_n I + Q_n), \qquad \operatorname{tr}(Q_n) = 0. \tag{6}$$

We call this representation *Denoising Strain* and propose Q_n as its tensor order parameter. The scalar m_n records common scale change; Q_n records the directional differences; R_n records transport. Mathematically, H_n is Hencky strain and Q_n its deviatoric part (Neff et al., 2016); we repurpose the latter as a generative order parameter. In the continuous limit, m_n corresponds to integrated divergence per dimension (Song et al., 2021). An isotropic transition has $Q_n = 0$ regardless of whether it expands or contracts. When the spectrum splits, Q_n acquires principal axes. Writing

$$\rho_n = \|Q_n\|_{\mathrm{F}}, \qquad \widehat{Q}_n = Q_n / \rho_n \tag{7}$$

separates anisotropy strength from orientation and spectral shape when $\rho_n > 0$. Here $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, the square root of the sum of squared matrix entries; at exact isotropy, $Q_n = 0$ but $\widehat{Q}_n$ is undefined. The actual stability of axis i remains $h_{n,i} = m_n + \eta_{n,i}$, where $\eta_{n,i}$ is the corresponding eigenvalue of Q_n.

We use the Elucidating the Design Space of Diffusion-Based Generative Models (EDM) framework and sampler of Karras et al. (2022). Let σ_n denote the noise level at state n. Setting the denoiser output to zero gives a reference that treats all directions equally; we call this the zero-denoiser baseline. Before the terminal zero-noise transition, its step is $x_{n+1} = (\sigma_{n+1}/\sigma_n) x_n$, hence $H_n = \log(\sigma_{n+1}/\sigma_n) I$ and $Q_n = 0$. It shrinks every direction equally and is invariant under every rotation of the input tangent space: no axis is preferred. A nonzero Q_n breaks this local rotational symmetry by selecting eigenspaces. This makes Q_n an order parameter for directional organization. Noise normalization, x_n/σ_n, leaves Q_n unchanged: it adds the isotropic term $\log(\sigma_n/\sigma_{n+1}) I$ to H_n (Appendix E).

2.3 Following directions through denoising

Stretch accumulated over several steps depends on how their axes align: transport can carry an expanding direction into one that contracts at the next step. We therefore track both incoming and outgoing axes. The eigenvectors of Q_n, which are also those of H_n, first give coordinates for local trajectory differences. For a displacement from the reference trajectory, we write

$$r_{n,i} = v_{n,i}^\top \delta x_n. \tag{8}$$

The sign of $r_{n,i}$ distinguishes the two sides of an axis. Flipping an eigenvector's sign simply exchanges the two intervention labels. Crossing $h_{n,i} = 0$ marks the *local expansion boundary*: the sampler changes from suppressing differences along that axis to amplifying them.

The output principal axis of a nonzero singular direction is

$$u_{n,i} = \frac{T_n v_{n,i}}{\exp(h_{n,i})}. \tag{9}$$

Individual eigenvectors can exchange order or become ambiguous when eigenvalues are close. We therefore compare the expanding directions collectively, weighting each by its positive log stretch:

$$P^{+}_{n,\mathrm{out}} = \sum_i [h_{n,i}]_+ u_{n,i} u_{n,i}^\top, \qquad P^{+}_{n+1,\mathrm{in}} = \sum_j [h_{n+1,j}]_+ v_{n+1,j} v_{n+1,j}^\top. \tag{10}$$

Here $[a]_+ = \max(a, 0)$. Write $A = P^{+}_{n,\mathrm{out}}$ and $B = P^{+}_{n+1,\mathrm{in}}$. Their normalized Frobenius inner product, $C_n = \langle A, B\rangle_{\mathrm{F}}/(\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}})$, measures alignment between the outgoing and next incoming expanding directions. For nonzero tensors, it equals one when they are proportional: their axes and relative positive stretches agree. It is invariant to eigenvector signs and orderings. Repeating these measurements along a sample path gives its *denoising-strain trajectory*: how directional organization, absolute expansion, and transport evolve throughout generation.

2.4 Leading-spectrum computation and checkpoint comparison

We estimate the k leading eigenpairs of H_n without materializing an image-space matrix. Differences between their eigenvalues measure directional organization without estimating the full-spectrum mean m_n. At each chosen state, the measurement follows three steps:

1. Differentiate the complete sampler step. Apply the Gram operator $A_n = T_n^\top T_n$ through $v \mapsto T_n^\top (T_n v)$, with a Jacobian–vector product followed by a vector–Jacobian product, keeping network weights fixed.
2. Run Lanczos on this operator with full reorthogonalization. Retain the k largest Gram eigenpairs (μ_i, v_i) and check their residuals.
3. Read off log stretches $h_i = \frac{1}{2}\log \mu_i$ and output axes $u_i = T_n v_i / e^{h_i}$. Repeat along the trajectory or at a new checkpoint.

Appendix B reports residuals, retained-rank sensitivity, and cost. To detect expansion, we test whether $h_i > 0$ before subtracting any common scale. The spectral spread $s_n = h_1 - h_k$, ordered $h_1 \geq \cdots \geq h_k$, removes the unknown common shift m_n and measures directional differentiation. Within the k computed eigenpairs, we summarize the number of active expanding directions by the participation ratio $d_n^+ = (\sum_i [h_{n,i}]_+)^2 / \sum_i [h_{n,i}]_+^2$. It equals the number of directions when their positive stretches are equal and decreases when expansion is concentrated in fewer axes. We report zero when no retained direction expands.

The spread describes how much the retained directions differ in stretch. To compare their orientation and strength across checkpoints, we construct a tensor from the same eigenpairs after subtracting the smallest retained eigenvalue. We call these k directions the retained sector and compare its tensor with the final-checkpoint reference. Similarity is one when both strength and orientation match, and near zero for negligible retained anisotropy. This measures how the directional geometry develops toward its final state (Appendix A).

3 Do the measured directions affect generation?

We test the geometry by perturbing its axes on pretrained animal-face and human-face models. Local-gain checks validate the measurement; completed images and held-out feature-space tests reveal its generative effect.

3.1 Experimental setup

We test the same measurement and intervention procedure on the unconditional 64×64 AFHQv2 and FFHQ models released with EDM (Karras et al., 2022), with dataset provenance from Choi et al. (2020) and Karras et al. (2019). For each dataset, we sample ten fixed latent trajectories with the deterministic 40-step EDM Heun sampler and compute the leading eight eigenpairs of each finite-step Gram operator at denoising states 10–18. Appendix B reports sensitivity to retaining $k = 4, 6, 8$ directions.

For causal validation, we select the state with maximum leading log stretch on each trajectory. From the same state x_n, we continue sampling after applying $+\epsilon v_1$ and $-\epsilon v_1$, where v_1 is the leading strain axis and the injected perturbation has root-mean-square (RMS) magnitude $0.01\sigma_n$ per state dimension. For each of four random control directions, we remove the component along v_1 and rescale to

the same perturbation norm. We validate one-step log gain with a separate $0.0005\sigma_n$ local probe, and measure downstream endpoint effect by the pixel-space RMS distance between the two completed images. A second protocol tests effects across denoising states using the continuous denoising flow. With time oriented toward decreasing noise, $t = -\sigma$, its velocity is $f_t(x) = (D_\theta(x, \sigma) - x)/\sigma$, where D_θ is the denoiser. The symmetric Jacobian $S_t = (D_x f_t + (D_x f_t)^\top)/2$ gives the local strain rate; over a small increment $\Delta t = \sigma_n - \sigma_{n+1} > 0$, the strain is approximately $\Delta t\, S_t$. We follow a frame of four orthonormal axes within the eight leading directions of this geometry on 200 trajectories per dataset. Its coordinates are projections of the noise-normalized state x_n/σ_n onto these axes. A linear predictor fitted on 100 trajectories maps these coordinates to features of the completed image; we call this a feature readout. We evaluate signed interventions on the other 100 trajectories at every state 8–18 and at three perturbation magnitudes. Appendix C defines the coordinates, features, and controls.

3.2 Directions begin to expand as denoising proceeds

The measured trajectories begin with contraction and develop expanding directions later in denoising (Figure 2). At state 10, the median leading log stretch is -0.046 on AFHQv2 and -0.036 on FFHQ. The leading direction contracts on all ten AFHQv2 trajectories and nine of ten FFHQ trajectories. The median leading log stretch becomes positive at state 12 on AFHQv2 and state 13 on FFHQ, then reaches 0.366 and 0.170, respectively, at state 18. More directions expand over the same interval: the participation ratio d_n^+ reaches 3.22 on AFHQv2 and 3.54 on FFHQ at state 18.

The leading subspace rotates across steps but remains strongly aligned after transport: median C_n is 0.865 on AFHQv2 and 0.889 on FFHQ. We compute this alignment for adjacent steps that both have expanding directions (Equation 10). Denoising thus carries its preferred directions forward while new directions become expanding.

Figure 2 also compares the measured finite-step strain with its infinitesimal approximation $\Delta t\, S_t$ defined above. The broad trends agree, but AFHQv2's median expansion boundary occurs at state 12 for the executed step and state 13 for the approximation (Appendix B.1). The infinitesimal approximation can therefore change the contraction/expansion reading.

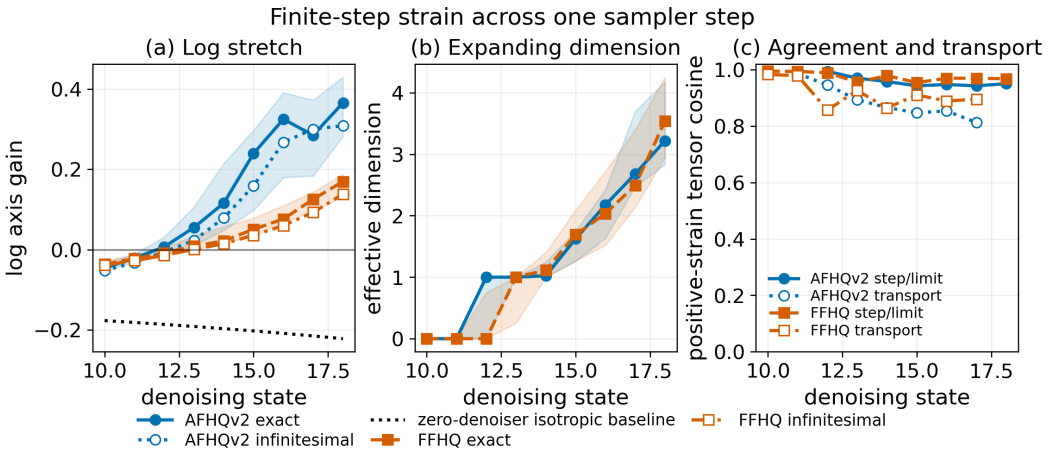


Figure 2: **Finite-step strain resolves local expansion and frame transport.** Left: the leading exact log stretch crosses zero during denoising and closely follows its infinitesimal approximation. Middle: the effective dimension of the positive measured spectrum grows. Right: "step/limit" compares the positive-strain tensors of the exact step and its infinitesimal approximation at the same state; "transport" compares the outgoing tensor of one step with the incoming tensor of the next. Both use normalized Frobenius inner products. Curves show medians over ten trajectories; bands show interquartile ranges.

3.3 SMALL INTERVENTIONS LEAD TO DIFFERENT COMPLETED IMAGES

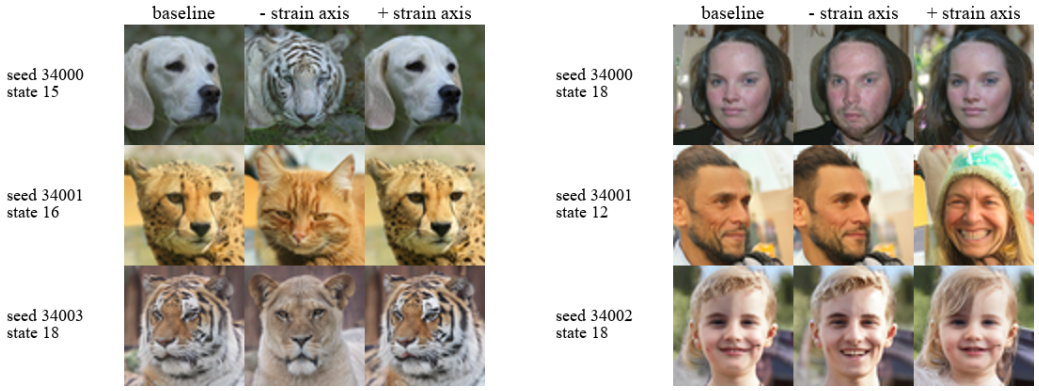
Perturbing the leading strain axis changes the final image substantially more than perturbing an equal-norm random direction. The median target-to-control endpoint separation is $27.56\times$ on AFHQv2 and $24.40\times$ on FFHQ; the target exceeds the median control on all 20 trajectories (Table 1). Thus the sensitivity measured at an intermediate state leads to visible differences in the completed images.

We also check that the computed strain predicts the immediate perturbation gain. The prediction is the leading eigenvalue $h_{n,1}$; the observed log gain is $\log(\|\delta x_{n+1}\|/\|\delta x_n\|)$. Their median absolute difference is 0.00576 on AFHQv2 and 0.000326 on FFHQ. This check validates the local measurement independently of its downstream effect.

Table 1: **Signed leading-axis intervention.** Endpoint ratio compares the separation induced by the discovered axis with the median separation from four equal-norm random controls orthogonal to the target axis. A win means that the target separation exceeds this median control on one trajectory.

Dataset	Median abs. log-gain error ↓	Endpoint ratio ↑	Wins
AFHQv2	0.00576	27.56×	10/10
FFHQ	0.000326	24.40×	10/10

The endpoint galleries show what these separations mean (Figure 3). AFHQv2 axes switch animal categories or alter breed and pose; FFHQ axes change identity, apparent age, expression, gender presentation, and viewpoint. These local alternatives are identified from the intermediate sampler state, without semantic labels or inspection of either endpoint.



(a) AFHQv2 (b) FFHQ

Figure 3: **Leading-axis perturbations change the generated image.** Three representative trajectories per dataset are shown. Each row starts from the same intermediate state and compares the baseline continuation with negative and positive perturbations. No image label is used to find the axis; Table 1 reports quantitative results over all ten trajectories per dataset.

A separate, larger experiment finds stronger feature-space responses along the tracked axes than along random controls at all 11 denoising states on both datasets. It tests the continuous-step counterpart of the finite-step geometry on 100 held-out trajectories per dataset. We encode completed images with the frozen DINOv2 image-feature model (Oquab et al., 2024). The signed response projects the feature difference between positive and negative interventions onto the direction predicted by the readout. Its dose-response slope is higher for tracked axes, with a 95% bootstrap interval above zero for the paired slope difference at every state. At the largest dose, the pixel effect is also 22.4–$34.0\times$ control on AFHQv2 and 18.3–$31.7\times$ on FFHQ. Appendix C also reports the signed-response tests at each dose.

The tracked coordinates also predict features from a second frozen image encoder, CLIP (Radford et al., 2021), on unperturbed held-out images at all 11 AFHQv2 states and nine of 11 FFHQ states (Appendix C).

4 EARLY GEOMETRIC ORGANIZATION DURING TRAINING

When does this generative organization form during training? MNIST provides a detailed view across noise levels; CIFAR-10 relates its emergence to FID. At each checkpoint, we measure the intermediate states produced by that model from fixed latent seeds and a fixed noise schedule. We follow development both across training exposure and along denoising.

4.1 CHECKPOINTS FROM UNINTERRUPTED TRAINING RUNS

The 32×32 MNIST EDM chain (LeCun et al., 1998) has ten snapshots at 0.256, 5, 10, 30, 100, 300, 1000, 3000, 10000, and 30000 kimg. All snapshots share one optimizer and exponential-moving-average (EMA) history; one kimg denotes 1000 training images. We measure six trajectories over states 8–30 of a 40-step deterministic Heun sampler. The unconditional CIFAR-10 chain uses EDM's DDPM++ denoiser architecture (Krizhevsky, 2009; Karras et al., 2022). It is continuously trained from initialization to 200000 kimg, with 18 checkpoints spaced approximately uniformly in log exposure. We measure eight trajectories across all 18 Heun transitions, using 32 Lanczos iterations per transition to retain eight leading Gram eigenpairs.

MNIST uses feature-FID from an independent digit classifier; CIFAR-10 uses standard FID. We compare geometric events with the *half-log-FID checkpoint*, which completes half the decrease in log FID from initialization to the best evaluated checkpoint. Appendix D.1 gives evaluator and event definitions.

We identify geometric events with criteria fixed before the full evaluations. At each checkpoint, we take the median across seeds at each state, then the maximum across states. Expansion is detected when this leading log stretch is positive at two consecutive checkpoints. The anisotropy threshold is a spectral spread of 0.001. On CIFAR-10, we additionally require tenfold growth from the near-initial spread and persistence at the next checkpoint.

4.2 FROM NEARLY EQUAL STRETCHES TO PREFERRED DIRECTIONS

The first sign of organization is a separation of previously near-equal stretches. On MNIST, the peak spectral spread is 6.0×10^{-7} at 0.256 kimg and every measured strain eigenvalue is negative. Leading-sector anisotropy appears before a measured axis expands: the spread reaches 0.00626 at 30 kimg, while the first positive median leading strain occurs at 100 kimg. The main reorganization occurs between 100 and 1000 kimg, when recognizable digits appear and retained-sector similarity to the 30,000-kimg reference rises from 0.067 to 0.651.

Higher-noise geometry matures earlier. At 3000 kimg, similarity to the final checkpoint is about 0.80–0.92 at states 8–20, but only 0.03–0.07 at states 28 and 30. Later checkpoints continue to reorganize the low-noise end of the trajectory (Appendix D).

Across Euler and Heun. A four-checkpoint MNIST comparison shows the same formation pattern: the solvers agree on the median leading-strain sign at all 16 checkpoint–state combinations, using six paired trajectories. Magnitudes remain sampler-specific: at 300 kimg and state 24, median leading strain is 0.115 for Heun and 0.195 for Euler. Appendix B.2 gives the complete paired results.

4.3 GEOMETRIC EVENTS PRECEDE THE FID PROGRESS LANDMARK

Geometric organization precedes substantial FID progress on both datasets. On MNIST, the spread threshold is crossed at 30 kimg, persistent expansion begins at 100 kimg, and half-log-FID arrives at 300 kimg.

On CIFAR-10, persistent expansion first occurs at 10 kimg, and the spectral spread crosses its threshold at 20 kimg (Figure 4). A separate measure of anisotropy, the Euclidean norm of the retained log stretches after subtracting their mean (Appendix D.1), has already grown $422\times$ from initialization at 10 kimg, while FID has completed only 0.0348% of its total log-scale improvement. The half-log-FID checkpoint arrives at 5000 kimg, requiring 500 times the exposure of the positive-strain event and 250 times that of the spectral-spread event.

The events distinguish directional organization from expansion. MNIST crosses the spread threshold first; CIFAR-10 develops a weak expanding axis first. The decomposition explains why: Q_n measures directional differences, while $H_n = m_n I + Q_n$ also includes common scale. Differentiated directions need not yet amplify perturbations.

Table 2 summarizes timing across alternative spread thresholds and expansion margins. Every geometric event in this grid precedes all three FID landmarks. Even the latest CIFAR-10 event precedes the earliest landmark by $25\times$. Raising the MNIST spread threshold further to 0.1 moves its crossing to the half-log-FID checkpoint (Appendix D.2). The spread-event times remain at 30/20 kimg on MNIST/CIFAR-10 when the stored spectral window is truncated from eight axes to six or four (Appendix D.3).

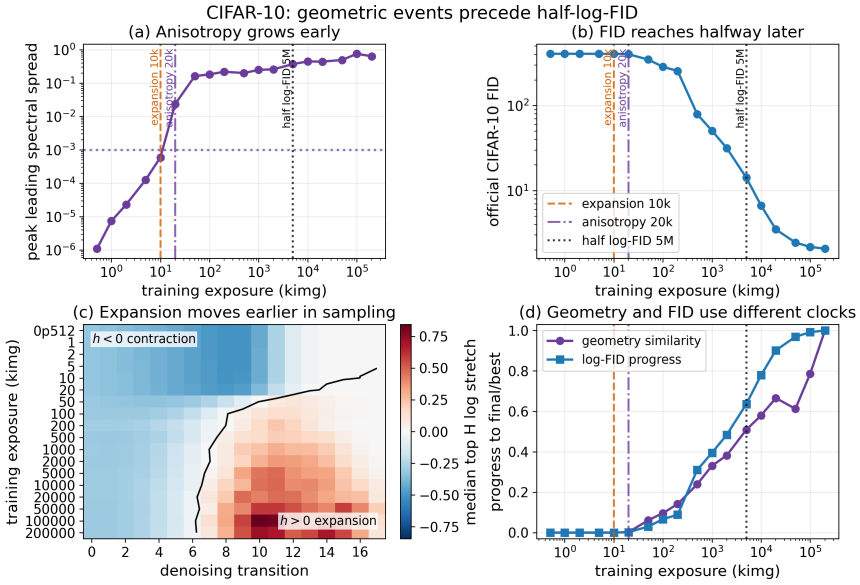


Figure 4: **Early geometric organization on CIFAR-10.** Leading-sector anisotropy and positive strain appear early, then extend toward earlier denoising transitions. Official 50,000-sample FID reaches its half-log-improvement checkpoint later, when retained-sector reference similarity is roughly halfway to its terminal value. Panels (a–c) locate the onset of structure; panel (d) follows its subsequent maturation toward the final geometry. The anisotropy curve in panel (a) is the maximum over denoising states of the seed-median leading spectral spread, the quantity used for the spread event.

Table 2: **Geometric events precede FID landmarks across alternative criteria.** Post hoc sensitivity on the same runs; all exposures are in kimg. Spread uses $0.1/0.3/1/3/10$ times the original threshold; expansion uses margins $0/10^{-6}/10^{-5}/10^{-4}/10^{-3}$. Ranges cover all tested settings.

Dataset	Spread event	Expansion event	FID 25% / 50% / 75%
MNIST	10–100	100	300 / 300 / 1000
CIFAR-10	5–20	10–20	500 / 5000 / 10000

Geometry and output metrics diverge late in MNIST training. Between 3000 and 10000 kimg, feature-FID worsens from 5.99 to 11.78 and recall falls slightly, while reference similarity rises from 0.804 to 0.870. Held-out loss, precision, and image morphology also improve (Appendix D).

5 Related Work

Evaluation of generative models. FID compares feature distributions (Heusel et al., 2017), while precision and recall separate quality from coverage (Kynkäänniemi et al., 2019). Finite-sample FID

can carry model-dependent bias (Chong & Forsyth, 2020). Our sampler-level geometry complements these output measurements by following how the generative process forms.

Diffusion dynamics and phase transitions. Diffusion learns reverse-time dynamics through denoising or score matching (Ho et al., 2020; Song et al., 2021); EDM distinguishes preconditioning, schedules, and solvers (Karras et al., 2022). Prior work derives symmetry breaking through fixed-point stability (Raya & Ambrogioni, 2023) and thermodynamic order parameters (Ambrogioni, 2025), describes separation toward data modes and individual samples (Biroli et al., 2024), and connects these transitions to data-manifold geometry (Yaguchi et al., 2025). Our tensor resolves local directions along trajectories of the executed finite sampler map, and we track how these directions develop across training checkpoints.

Spectral geometry of the score and sampler. The closest line of work studies spectra of Jacobians of local diffusion dynamics. Score-Jacobian gaps reveal geometric phases (Ventura et al., 2025), while eigenvalues of the symmetric drift Jacobian quantify local contraction and expansion (Stančević & Ambrogioni, 2026). Scores also reveal intrinsic dimension (Stanczuk et al., 2024). Recent preprints use the score-Jacobian Gram matrix as a Riemannian metric for manifold-aware paths (Saito & Matsubara, 2025), posterior covariance spectra for out-of-distribution detection (Shoushtari et al., 2025), and singular geometric events to locate intervention-sensitive sampling intervals (Sakamoto & Sakamoto, 2026). Denoising Strain measures how the executed sampler step deforms perturbations. Its traceless log strain isolates anisotropy; we test the generative effect of its axes and follow its formation through training.

6 Discussion and future work

Denoising Strain follows how a sampler develops preferred directions during training. On pretrained models, interventions along the strain axes change completed images, linking local stretching to the model's generative behavior. This directional structure appears early in training: on CIFAR-10, geometric events precede the half-log-FID checkpoint by 250–500 times in training exposure.

Future work. Our training evidence covers one run per dataset and a leading spectral window. Replicated runs at larger model and dataset scales would test generality and predictive value. Wider spectra, stochastic samplers, and feature-space distances would extend the observable. Diffusion Transformers provide a direct next setting; language models require defining states and transitions across layers. Measuring how candidate examples change Q and H could guide data selection, curricula, and update size. The goal would be to reduce the compute needed to reach a target performance with ordinary gradient-based optimization.

AI USE STATEMENT

The central research question, hypothesis, and scientific ideas in this work were conceived by the human author. Generative AI tools were used to automate the execution and bookkeeping of author-directed experiments, assist with literature search, and support manuscript drafting and editing, figure preparation and inspection, and paper review. All AI-assisted code, experimental outputs, citations, figures, and text were inspected and verified by the author. The author made the final scientific judgments and takes full responsibility for the content of the paper.

REFERENCES

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking, and critical instability. *Entropy*, 27(3):291, 2025. doi: 10.3390/e27030291. URL https://doi.org/10.3390/e27030291.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. doi: 10.1038/s41467-024-54281-3. URL https://doi.org/10.1038/s41467-024-54281-3.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Choi_StarGAN_v2_Diverse_Image_Synthesis_for_Multiple_Domains_CVPR_2020_paper.html.

Min Jin Chong and David Forsyth. Effectively unbiased FID and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.

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 *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1926. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Patrizio Neff, Bernhard Eidel, and Robert J. Martin. Geometry of logarithmic strain measures in solid mechanics. *Archive for Rational Mechanics and Analysis*, 222(2):507–572, 2016. doi: 10.1007/s00205-016-1007-x. URL https://doi.org/10.1007/s00205-016-1007-x.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. 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, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d0da30e312b75a3fffd9e9191f8bc1b0-Abstract-Conference.html.

Shinnosuke Saito and Takashi Matsubara. Be tangential to manifold: Discovering riemannian metric for diffusion models, 2025. URL https://arxiv.org/abs/2510.05509v3. Preprint, revised July 2026.

Ryosuke Sakamoto and Kotaro Sakamoto. The geometry of phase transitions in generative dynamics via projection caustics, 2026. URL https://arxiv.org/abs/2606.13191. Preprint.

Shirin Shoushtari, Yi Wang, Xiao Shi, M. Salman Asif, and Ulugbek S. Kamilov. EigenScore: OOD detection using covariance in diffusion models, 2025. URL https://arxiv.org/abs/2510.07206v1. Preprint.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Dejan Stančević and Luca Ambrogioni. The information dynamics of generative diffusion. *Entropy*, 28(2):195, 2026. doi: 10.3390/e28020195. URL https://doi.org/10.3390/e28020195.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440, 2024. URL https://proceedings.mlr.press/v235/stanczuk24a.html.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dbdb2729f874e209b56ffecd23aa3b36-Abstract-Conference.html.

Manato Yaguchi, Kotaro Sakamoto, Ryosuke Sakamoto, Masato Tanabe, Masatomo Akagawa, Yusuke Hayashi, Masahiro Suzuki, and Yutaka Matsuo. The geometry of phase transitions in diffusion models: Tubular neighbourhoods and singularities. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=ahVFKFLYk2.

A Comparing geometry across training checkpoints

For an initial latent seed z, we collect the strain, anisotropy, and transport along its sampling trajectory:

$$\mathcal{S}_\theta(z) = \{H_{\theta,n}(x_n(z)), Q_{\theta,n}(x_n(z)), R_{\theta,n}(x_n(z))\}_{n=0}^{N-1}. \tag{11}$$

Across checkpoints we compare the paths from the same latent seeds.

To compare checkpoints, let c index checkpoints and $\star$ denote the terminal evaluated reference. From the k computed eigenpairs, we construct the scalar-shift-invariant retained tensor

$$\widetilde{Q}_{c,n,z} = \sum_{i=1}^{k} (h_{c,n,z,i} - h_{c,n,z,k}) v_{c,n,z,i} v_{c,n,z,i}^\top, \tag{12}$$

Its magnitude-sensitive similarity to the reference is

$$M_{c,n,z} = \frac{2\langle \widetilde{Q}_{c,n,z}, \widetilde{Q}_{\star,n,z} \rangle_{\mathrm{F}}}{\left\| \widetilde{Q}_{c,n,z} \right\|_{\mathrm{F}}^2 + \left\| \widetilde{Q}_{\star,n,z} \right\|_{\mathrm{F}}^2}. \tag{13}$$

Similarity is one when the retained tensor matches the reference in both strength and orientation. To see how it behaves near isotropy, write $A = \widetilde{Q}_{c,n,z}$, $B = \widetilde{Q}_{\star,n,z} \neq 0$, and $r = \|A\|_{\mathrm{F}} / \|B\|_{\mathrm{F}}$. Both tensors are positive semidefinite, so Cauchy–Schwarz gives

$$0 \leq M_{c,n,z} \leq \frac{2r}{1+r^2}. \tag{14}$$

Thus a vanishing retained anisotropy has vanishing similarity regardless of its axes. Rotations within an exactly degenerate eigenspace leave its weighted projector unchanged; at the retained boundary the weight is zero. The weighted tensor therefore remains well-defined within exact degeneracies. Appendix B examines numerical accuracy and the retained window.

B Numerical checks and comparisons of the measured dynamics

We estimate the leading eigenpairs of the executed sampler-step Gram operator as Ritz pairs from matrix-free Lanczos. Full reorthogonalization avoids duplicate Ritz values in clustered spectra. We pair every reconstructed vector with its Rayleigh quotient. For a unit eigenvector estimate v_i of the Gram operator $A = T_n^\top T_n$ and eigenvalue estimate μ_i, the normalized residual is $\|Av_i - \mu_i v_i\|_2 / \max(1, |\mu_i|)$. It measures the error in the eigenvalue equation while retaining absolute error control near zero. The maximum over the reported AFHQv2 and FFHQ finite-step scans is 9.78×10^{-4}. Near an isotropic operator, we record early Krylov breakdown as degeneracy.

Table 3 recomputes the summary statistics from the already stored top-eight eigenpairs. The leading-axis crossings and interventions are unchanged by construction. Transport alignment is unchanged to three decimals between $k = 6$ and $k = 8$, and the state-18 participation ratio is nearly converged by $k = 6$. Increasing k widens the observed spectral window.

Table 3: **Sensitivity to the retained eigenpair budget.** Alignment is the median C_n over valid seed–step pairs; d_{18}^{+} is the seed-median participation ratio at state 18.

Dataset	Statistic	$k = 4$	$k = 6$	$k = 8$
AFHQv2	transport alignment	0.861	0.865	0.865
	d_{18}^{+}	2.675	3.192	3.218
FFHQ	transport alignment	0.889	0.889	0.889
	d_{18}^{+}	3.149	3.537	3.537

On the hardware used for the finite-step scans, nine measured states require a median 353 seconds per AFHQv2 trajectory and 366 seconds per FFHQ trajectory, with 3.67 GiB peak allocated GPU memory. States and trajectories are independent and parallelize directly. For L Lanczos iterations and k retained axes, one state uses L Jacobian–vector products (JVPs, applying T_n) and L vector–Jacobian products (VJPs, applying $T_n^{\top}$) for the Gram iteration, another k of each for Rayleigh-quotient and residual checks, and k JVPs to construct output axes: $L + 2k$ JVPs and $L + k$ VJPs in total. Parameter gradients are never accumulated.

B.1 Where finite-step and infinitesimal readings differ

We compare the largest measured finite-step log stretch with the largest eigenvalue of $\Delta t\, S_t$ at every stored state on all ten trajectories per pretrained model. This post hoc paired reanalysis uses the same model, state, and step interval for both readings. Across 90 state–trajectory pairs per dataset, the median absolute difference in leading log stretch is 0.02296 on AFHQv2 and 0.01153 on FFHQ. The estimated signs disagree on 8/90 and 5/90 pairs, respectively. Each count pools nine states from each of ten trajectories.

The leading directions usually agree strongly across this entire window. Writing v and u for the unit leading axes of the finite-step and infinitesimal operators at the same state, the mean squared overlap $(v^{\top}u)^2$ over all 90 pairs is 0.9413 on AFHQv2 and 0.9579 on FFHQ; the medians are 0.9605 and 0.9776. Individual overlaps reach as low as 0.5803 and 0.6154.

On AFHQv2 at state 12, the seed-median leading log stretch is $+0.00649$ for the executed step but -0.00595 for the approximation. The median squared overlap of their leading axes is 0.9906: nearly aligned directions can still receive different contraction/expansion readings. The first positive median is at state 12 for the exact step and 13 for the approximation; on FFHQ both occur at state 13.

B.2 Changing the executed sampler: Euler and Heun

The qualitative formation pattern recurs with both solvers, while its magnitude depends on the executed dynamics (Table 4). At 30 kimg, the stretches have separated but all four seed-median leading strains remain negative. At 100 kimg, the median is positive at state 30; at 300 kimg, it is also positive at the earlier state 24. The two solvers agree on the median sign at all 16 checkpoint–state combinations, with one sign disagreement among the 96 individual paired measurements.

We fixed four existing MNIST EMA checkpoints (10, 30, 100, 300 kimg), six paired CPU latent draws (seeds 42000–42005), and states 8, 16, 24, 30 before comparing solvers. Both use the same 40-step EDM noise schedule and initial latent; each follows its own trajectory. We differentiate the complete Euler or Heun update, including Heun's corrector, retaining eight Gram eigenpairs. The sampler uses float64 states and the derivative input is float32. These CPU draws differ from the original CUDA scans. Heun's corrector adds denoiser evaluations at the same step count.

The calculation starts with 32 Lanczos iterations. The paired refinement rule recomputes both solvers at 64 iterations wherever either normalized Gram residual exceeds 10^{-3}, then at 128 if needed. We fixed this rule after the initial calculation and before refinement. Nine paired points passed at 64 and one at 128; the remaining 86 needed no refinement. The final maximum residual is 8.00×10^{-4}. Refinement changes individual leading log stretches by at most 6.92×10^{-6} and retained spreads by at most 0.01587; the table uses the refined values.

Table 4: **A shared formation pattern, different stretch magnitudes.** Six-seed medians for every prespecified checkpoint and state. h_1 is the leading log stretch and $s = h_1 - h_8$ the retained spectral spread. Noise levels at states 8/16/24/30 are approximately 23.61/5.38/0.823/0.131.

kimg	State	h_1 Heun	h_1 Euler	s Heun	s Euler
10	8	-0.1686	-0.1686	0.000003	0.000003
	16	-0.2066	-0.2069	0.000015	0.000015
	24	-0.1823	-0.1922	0.000085	0.000088
	30	-0.0164	-0.0193	0.000050	0.000055
30	8	-0.1681	-0.1681	0.000194	0.000194
	16	-0.2042	-0.2045	0.001074	0.001073
	24	-0.1685	-0.1780	0.006322	0.006459
	30	-0.0086	-0.0108	0.003598	0.003901
100	8	-0.1640	-0.1641	0.002477	0.002469
	16	-0.1820	-0.1825	0.013728	0.013378
	24	-0.0051	-0.0263	0.113316	0.101483
	30	$+0.0635$	$+0.0681$	0.051290	0.056753
300	8	-0.1512	-0.1516	0.012335	0.011958
	16	-0.0568	-0.0780	0.109589	0.089969
	24	$+0.1150$	$+0.1951$	0.180734	0.195999
	30	$+0.0416$	$+0.0486$	0.021793	0.025285

C LARGE-SAMPLE VALIDATION OF THE MOVING FRAME

The continuous-time geometry lets us follow a moving frame on 200 AFHQv2 and 200 FFHQ trajectories. Using the generation-time convention $t = -\sigma$ and strain rate S_t defined in Section 3, a sufficiently small sampler step gives

$$H_n = \Delta t\, S_t + O(\Delta t^2), \tag{15}$$

so the finite-step strain and its continuous approximation share their leading geometry in the small-step limit. We tracked four axes within the eight leading directions computed at states 8–18.

We compare two evolving frames: one shared by all trajectories at each state, and one that follows each individual trajectory. The shared frame consists of the four leading eigenvectors of the mean projector onto local top-eight subspaces from a separate set of 20 trajectories. We compute it separately at each state, align these frames between states, and freeze the resulting sequence before the 200-trajectory evaluation. We construct the tracked frame within each trajectory's local eight-dimensional subspace, choosing four axes that maximize overlap with a reference frame. At the first state, the reference is the shared frame; thereafter, it is the preceding tracked frame.

For two four-axis frames, projector overlap is the mean squared singular value of their cross-product; the weakest frame cosine is its smallest singular value. They summarize, respectively, average subspace agreement and the least aligned direction. The tracked frame remains continuous while departing from the shared frame. At state 18, the weakest consecutive frame cosine is 0.801 on AFHQv2 and 0.821 on FFHQ. Eigenvalues of the continuous-step strain restricted to the tracked frame cross zero at different denoising states, matching the sequence of local contraction-to-expansion changes observed with the finite-step strain.

We use three sets of four coordinates to predict the unperturbed final image's DINOv2-small features. Each coordinate is the projection of the intermediate state, divided by its noise level, onto a unit frame axis: $b_i^\top x_n/\sigma_n$. The tracked coordinates use the trajectory-specific frame; the shared coordinates use the common frame at that state. The random-control coordinates use a fixed random orthonormal frame orthogonal to the shared frame. These coordinates project the noise-normalized state itself; $r_{n,i}$ in Equation 8 projects a displacement from the reference trajectory.

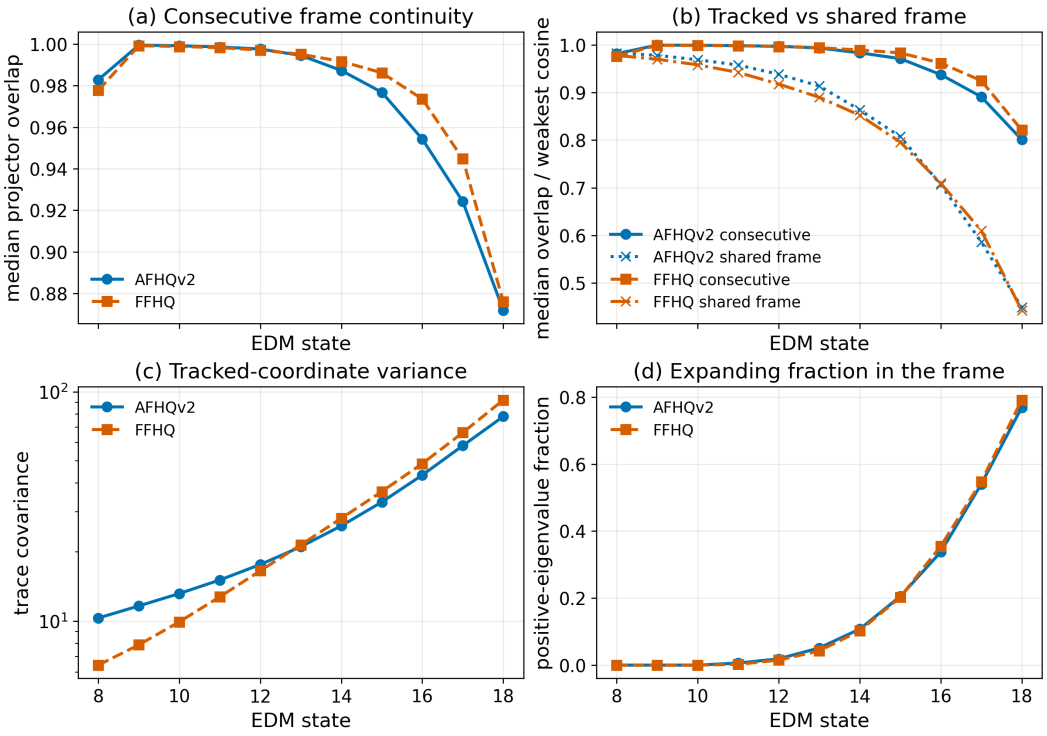


Figure 5: **Large-sample frame transport.** A trajectory-conditioned frame remains continuous across denoising time while rotating away from the state-specific shared frame. Panel (a) reports consecutive projector overlap; (b) compares the weakest consecutive frame cosine with shared-frame projector overlap. Panel (c) gives the summed variances of the noise-normalized state coordinates in the tracked frame; (d) gives the fraction of positive eigenvalues of the continuous-step strain restricted to that frame. Results aggregate 200 trajectories per dataset.

For each coordinate set and state, we fit a ridge regression on 100 trajectories and evaluate it on 100 held-out trajectories. Ridge regression uses an L2 penalty on the linear predictor's coefficients. Inputs are standardized using the training split. Let e_j be the unit-normalized DINOv2 feature vector for held-out image j, centered by the training-feature mean, and let $\widehat{e}_j$ be its prediction. The explained fraction is $1 - \sum_j \|e_j - \widehat{e}_j\|^2 / \sum_j \|e_j\|^2$. For the tracked coordinates, it ranges from 0.62% to 1.72% on AFHQv2 and 0.66% to 2.17% on FFHQ. Permutation tests use 1000 label permutations, with Benjamini–Hochberg false-discovery-rate (FDR) correction across the 11 states. Every tracked-coordinate readout attains the smallest possible adjusted q-value, whereas no random-control readout is significant. Shared-coordinate readouts also predict final features and achieve higher explained fractions than tracked-coordinate readouts (Figure 6).

Predicting a second visual representation. The coordinates also carry information about CLIP image features (Radford et al., 2021). We encode all 200 archived unperturbed endpoints per dataset with CLIP ViT-B/32, using its unit-normalized 512-dimensional image embeddings. We retain the original coordinates and 100/100 split, and fit new readouts with the same ridge and 1000-permutation procedure. Tracked-coordinate predictions pass the FDR test at all 11 AFHQv2 states and nine of 11 FFHQ states. The FFHQ tracked signal weakens at states 17–18; shared coordinates remain predictive at every state on both datasets, while random controls pass at none (Table 5). The endpoints are float16 archival copies, clipped and converted to 8-bit images before CLIP's standard preprocessing. Model revision and input hashes accompany the released analysis.

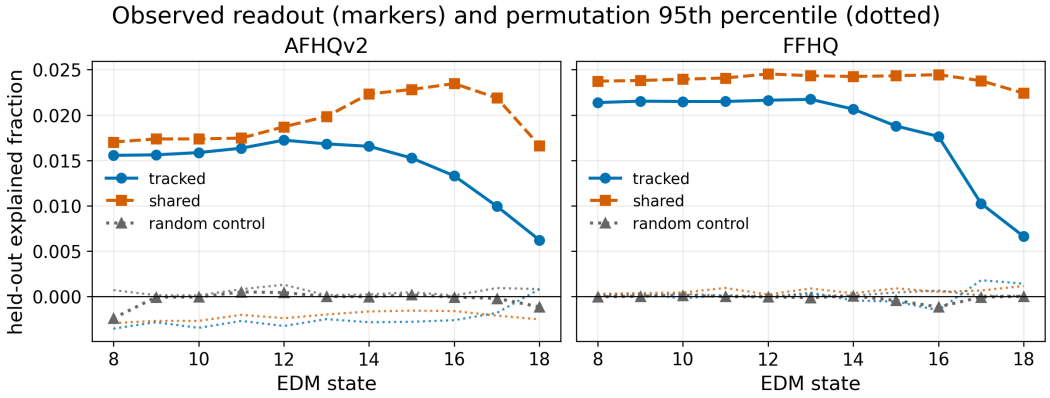


Figure 6: **The moving coordinates predict unperturbed final-image features.** "Tracked" uses a trajectory-specific frame; "shared" uses the frame common to all trajectories at that state; "random control" uses a frame orthogonal to the shared frame. Each curve reports a separate readout on the same frozen 100/100 train/test split. Dotted lines show the corresponding permutation-null 95th percentiles.

Table 5: **Geometry predicts a second visual representation.** Held-out CLIP explained fraction (%) at every measured state. A star marks permutation $q < 0.05$, corrected across the 11 states separately for each dataset and coordinate set.

	AFHQv2			FFHQ		
State	Tracked	Shared	Random	Tracked	Shared	Random
8	0.466^*	0.617^*	-0.022	0.360^*	0.495^*	-0.009
9	0.451^*	0.653^*	-0.013	0.343^*	0.487^*	-0.058
10	0.468^*	0.602^*	-0.009	0.318^*	0.484^*	0.047
11	0.525^*	0.625^*	0.014	0.285^*	0.480^*	0.005
12	0.628^*	0.711^*	0.006	0.408^*	0.483^*	-0.011
13	0.667^*	0.850^*	0.019	0.406^*	0.466^*	-0.005
14	0.656^*	1.121^*	-0.004	0.357^*	0.474^*	-0.021
15	0.544^*	1.160^*	0.001	0.216^*	0.518^*	-0.020
16	0.322^*	1.159^*	-0.007	0.116^*	0.596^*	-0.010
17	0.187^*	0.962^*	0.001	0.079	0.526^*	0.003
18	0.434^*	0.537^*	-0.564	0.003	0.435^*	0.002

Finally, signed interventions use RMS doses $0.00125\sigma_n$, $0.0025\sigma_n$, and $0.005\sigma_n$ at every measured state on the 100 held-out trajectories. For each tracked coordinate, its ridge-readout coefficient defines an axis in DINOv2 feature space. The signed response is the projection of the plus-minus final-feature difference onto this unit-normalized feature axis. We fit its slope through the three doses and compare it with the slope obtained from a matched orthogonal direction. Here the controls are random orthonormal directions outside the trajectory's local top-eight subspace; they differ from the shared-frame controls used for prediction above. The same readout feature axis is used to score target and control interventions. The median slope and median target-minus-control slope are positive at all 11 states on both datasets, with 95% bootstrap intervals above zero. Both decrease toward later states, indicating that equally scaled local interventions have progressively less influence on the final sample. At each dose, we also take the median signed response over the four axes for each trajectory and test whether more than half of these trajectory medians are positive. The one-sided binomial test passes FDR correction across states at every state and dose on both datasets.

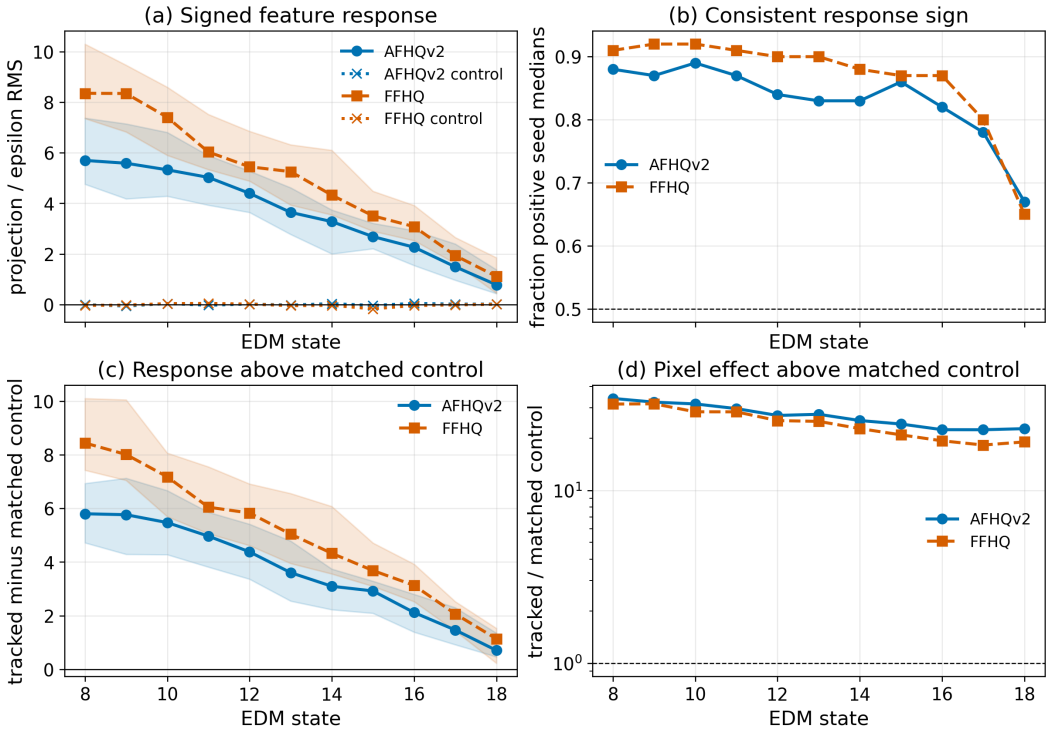


Figure 7: **Large-sample signed causal validation.** Perturbations along the tracked frame axes produce a dose-dependent final-feature response that exceeds matched orthogonal controls. "Tracked" denotes interventions along the trajectory-specific frame; controls are orthogonal to its local top-eight candidate subspace. Panel (a) shows slopes fitted through the three doses and the origin, summarized by taking the median over four axes within each trajectory, then over trajectories. Panel (b) gives the fraction of trajectories with a positive within-trajectory median slope; (c) shows the median paired difference between target and control slopes. Shading in (a) and (c) gives 95% bootstrap intervals. Panel (d) shows the median pixel-effect ratio over trajectory–axis pairs at the largest RMS dose, $0.005\sigma_n$.

D COMPLETE TRAINING-EVENT CURVES

MNIST develops its final-checkpoint geometry earlier at higher noise levels. At 3000 kimg, states 8/12/16/20 have reference similarities $0.917/0.917/0.882/0.796$, while states 24/28/30 have $0.385/0.068/0.028$. The reference is the 30,000-kimg checkpoint. Figure 9 shows how MNIST geometry, feature-FID, precision and recall (Heusel et al., 2017; Kynkäänniemi et al., 2019), morphology statistics, and held-out denoising loss change across checkpoints. Figure 4 shows the CIFAR-10 geometry and standard 50,000-sample FID curves.

D.1 DEFINITIONS OF THE TRAINING LANDMARKS

For each run, the half-log-FID checkpoint is the first checkpoint at which log FID has completed at least half of its total decrease from initialization to the best evaluated checkpoint. MNIST feature-FID uses the frozen 64-dimensional penultimate representation of an independent classifier trained only on the official training split; its test accuracy is 99.17%. CIFAR-10 uses the standard 50,000-sample FID evaluation.

The positive-leading-strain event is the first checkpoint at which the maximum over transitions of the median leading h_i is positive and remains positive at the next checkpoint. The qualifying transition need not be the same at both. The spread event uses the maximum over states of the seed-median spectral spread s_n. On MNIST it first reaches 0.001. On CIFAR-10 it first reaches the larger of 0.001 and ten times its near-initial value, and remains above that value at the next checkpoint. Table 6 lists the threshold-crossing checkpoints.

CIFAR-10 reaches persistent expansion at 10 kimg and the spread threshold at 20 kimg, reversing the ordering observed on MNIST. This also reverses the pre-evaluation protocol's prediction that the spread threshold would be crossed before expansion.

Table 6: **Training events occur on different clocks.** Events use criteria fixed before the corresponding full evaluation. Exposure is measured in kimg.

Dataset	Spread threshold	Positive leading strain	Half-log-FID
MNIST	30	100	300
CIFAR-10	20	10	5000

The $422\times$ growth on CIFAR-10 refers to the centered spectral magnitude: the Euclidean norm of the k retained log stretches after subtracting their mean. For each seed and state, we compute $\|(h_1 - \bar{h}, \ldots, h_k - \bar{h})\|_2$, where $\bar{h} = k^{-1} \sum_{i=1}^{k} h_i$, then report its median over seeds and states.

D.2 SENSITIVITY OF THE TRAINING-EVENT TIMING

To test how the detection criteria affect the timing lead, we reanalyze the stored checkpoint summaries, varying the spread threshold and FID landmark separately, using the same aggregation and persistence rules as the original analysis. Geometry occurs first in all 30 combinations (15 per dataset) in Table 7. We also test stronger expansion by raising the leading-strain threshold. Requiring persistent leading strain above 0, 10^{-6}, 10^{-5}, 10^{-4}, or 10^{-3} gives MNIST events at $100/100/100/100/100$ kimg and CIFAR-10 events at $10/10/10/20/20$ kimg. All 30 corresponding comparisons with the three FID landmarks also retain a lead.

Table 7: **Earlier geometry across alternative thresholds and FID landmarks.** Entries are first qualifying checkpoints in kimg. The original spread threshold is 0.001 on both runs; the original FID landmark is 50%.

Spread-threshold factor	0.1	0.3	1	3	10
MNIST	10	30	30	30	100
CIFAR-10	5	10	20	20	20
Log-FID progress	25%	50%	75%		
MNIST	300	300	1000		
CIFAR-10	500	5000	10000		

A wider absolute spread-threshold sweep of $10^{-6}, 10^{-5}, \ldots, 1$ gives MNIST crossings at $5/5/10/30/100/300$ kimg, with no crossing at 1; CIFAR-10 gives $0.512/2/5/20/20/50$ kimg, also with no crossing at 1. At the lowest threshold, CIFAR-10 already qualifies at the earliest checkpoint. At threshold 0.1, MNIST ties its half-log-FID landmark. Changing the threshold can also reverse the order of spread and expansion on CIFAR-10. The size of the timing lead therefore depends on which degree of geometric organization is being detected; Table 7 quantifies that dependence.

D.3 Training-event sensitivity to the retained window

The spread-event checkpoints are unchanged when the stored top-eight spectra are truncated to four or six axes. We compare $k = 4, 6, 8$ on all 60 MNIST scans (ten checkpoints, six seeds) and 144 CIFAR-10 scans (18 checkpoints, eight seeds). The $k = 8$ reconstruction exactly reproduces every state-wise seed-median spread and leading strain in the reported summaries. Holding the original 0.001 spread threshold and persistence rule fixed, all three windows give the same event checkpoints: 30 kimg on MNIST and 20 kimg on CIFAR-10.

For $k = 4, 6, 8$, the peak spreads immediately before the MNIST event are 0.000030, 0.000062, and 0.000106; at the event they are 0.004124, 0.005238, and 0.006261. The corresponding CIFAR-10 values are 0.000178, 0.000356, and 0.000598 before the event, and 0.016307, 0.021482, and 0.023708 at the event. Leading-strain events are unchanged by construction because every truncation retains the same largest eigenvalue.

D.4 Late-training metric disagreement on MNIST

The late-training disagreement between MNIST geometry and feature-FID persists in an independent evaluation. With 20,000 samples per checkpoint, feature-FID worsens from 5.985 at 3000 kimg to 11.782 at 10000 kimg, then improves to 4.576 at 30000 kimg. The paired bootstrap difference between the 10000- and 3000-kimg checkpoints is $+5.83$ with a 95% interval $[2.27, 9.33]$. Recall also decreases from 0.928 to 0.914 over the same interval.

Between 3000 and 10000 kimg, similarity to the final-checkpoint geometry rises from 0.804 to 0.870, precision rises from 0.752 to 0.795, the standardized discrepancy in classifier-independent digit morphology falls from 0.136 to 0.110, and held-out denoising loss falls from 0.1030 to 0.0980. The morphology statistic aggregates discrepancies in ink mass, foreground area, centroid, covariance shape, anisotropy, and total variation. Reference similarity is the median of $M_{c,n,z}$ over the six fixed seeds and measured states; held-out loss is averaged over 2048 fixed noise–image pairs.

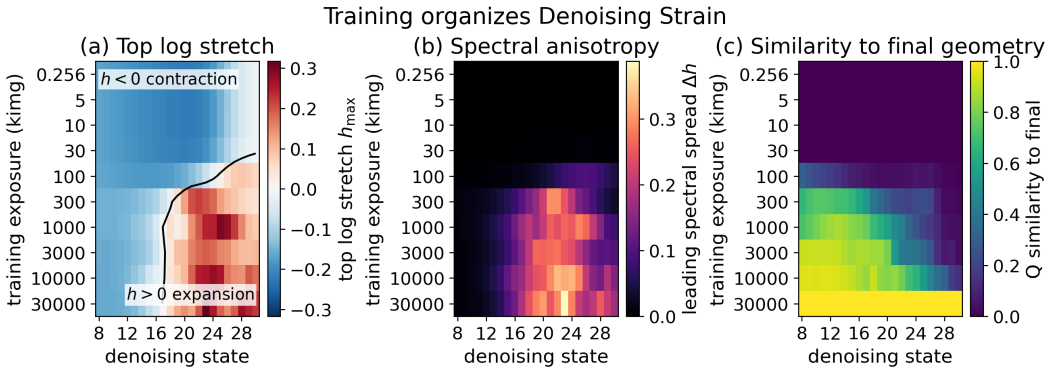


Figure 8: **Training organizes the measured leading strain sector on MNIST.** Left: positive leading strains appear across denoising states. Middle: the initially degenerate leading spectrum develops anisotropy. Right: k-dependent retained-sector similarity to the 30,000-kimg reference checkpoint.

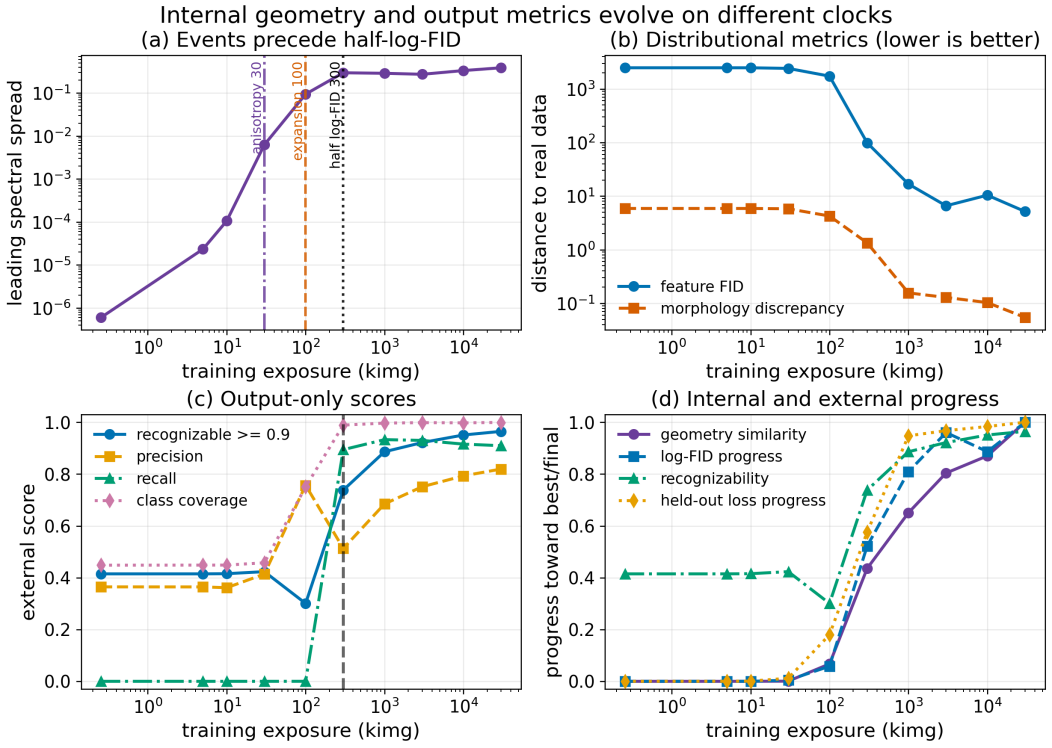


Figure 9: **Geometry and evaluation metrics across MNIST checkpoints.** The leading-spectrum spread criterion and the first persistent positive strain occur before the half-log-FID checkpoint. Later, retained-sector similarity to the final evaluated checkpoint continues to rise while feature-FID temporarily worsens. In panels (c) and (d), recognizability is the fraction of generated images whose highest digit-classifier probability is at least 0.9. Class coverage in panel (c) is the entropy of the predicted-label distribution, divided by $\log 10$; it is one for a uniform distribution over the ten digits.

E THE SAME DIRECTIONAL GEOMETRY IN DIFFERENT UNITS

The anisotropy tensor is unchanged by uniform changes of units at each state and rotates with an orthogonal change of coordinate axes. To see this, write $y_n = a_n O_n x_n$, where $a_n > 0$ and O_n is orthogonal. The tangent map in the new coordinates is

$$T_n^{(y)} = \frac{a_{n+1}}{a_n} O_{n+1} T_n O_n^\top. \tag{16}$$

With $c_n = a_{n+1}/a_n$, orthogonality gives $(T_n^{(y)})^\top T_n^{(y)} = c_n^2 O_n (T_n^\top T_n) O_n^\top$. Taking half the matrix logarithm and then subtracting its mean trace yields

$$H_n^{(y)} = O_n H_n O_n^\top + \log(c_n) I, \qquad Q_n^{(y)} = O_n Q_n O_n^\top. \tag{17}$$

Thus the eigenvalues of Q_n, its Frobenius norm, and the spectral spread are invariant; its axes transform by O_n. The common shift in H_n records the change in units between the two states.

For noise-normalized states $y_n = x_n/\sigma_n$ at positive noise levels, $O_n = I$ and $c_n = \sigma_n/\sigma_{n+1}$. Raw and noise-normalized states therefore give exactly the same Q_n and spread-event times. Every log stretch shifts by $\log(\sigma_n/\sigma_{n+1})$. In these units, the zero-denoiser reference is the identity map: $H_n^{(y)} = Q_n^{(y)} = 0$. Expansion then measures growth relative to the shrinking noise scale.