EVALUATING GENERATIVE MODELS VIA DIFFUSION ENERGY DIVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models now produce diverse outputs, including depth maps, point clouds, video, whose quality is naturally assessed with task-specific distances between individual samples, e.g. AbsRel for depth, Chamfer distance for point clouds. Turning these distances into divergences between distributions is difficult, though, as the standard tools like energy distance and distance-induced maximum mean discrepancy (MMD) require negative type, a global condition that several commonly used distances violate on real data. We introduce the Diffusion Energy Divergence (DED), which uses a task-specific distance only to connect samples to their nearest neighbors and weight the resulting graph edges. The normalized graph Laplacian is positive semidefinite for any symmetric, nonnegative weights, and its heat semigroup therefore yields a Euclidean diffusion embedding in which energy distance is a valid divergence. This embedding depends on the original distance only through the neighbor lists and edge weights. We fit the graph once on real data and place all evaluation and generated samples on it by a Nyström extension, giving every generator the same diffusion geometry. A permutation test combines the divergence with an off-support check and has exact size under the null for any such distance. On NYU depth maps, energy distance computed directly on AbsRel ranks two depth degradations in reverse order of severity (Spearman $\rho = -0.89$ for both), while DED ranks them correctly ($\rho = +0.96$ and $+1.00$). On VBench, DED detects mode collapse in 16 of 20 generators, although the collapse leaves the expected VBench scores unchanged.

1 INTRODUCTION

Generative models are increasingly being trained to produce diverse modalities such as depth maps (Bai et al., 2026), point clouds (Weijler et al., 2026), audio, video, and more recently the feature sets of the encoders used to score them (Zheng et al., 2025a). A generated sample has no ground truth to be compared against, so the per-sample measures that are used to measure performance do not work. Instead, evaluating a generator requires a comparison between distributions. We ask how we can turn such a per-sample measure into a divergence between distributions.

Energy distance (Székely & Rizzo, 2013) and MMD (Gretton et al., 2012) are the natural tools for this. Generalized energy distance takes a distance d directly, and MMD can take it through the distance-induced kernel of Sejdinovic et al. (2013). In both cases d must be of *negative type* (Lyons, 2013), a condition on its global behavior. Without it the energy distance can come out negative and the distance-induced kernel is no longer positive definite. (MMD with an arbitrary positive-definite kernel needs no such condition, but then the kernel, not a given distance choice, decides what counts as similar.) We measure negative type directly on diverse types of real data and find it violated by Chamfer distance on point clouds, AbsRel on depth, dynamic time warping on motion, best-RMSD alignment of molecular conformers, and Hausdorff distance between token sets (Table 15).

The key insight is that these distances are usually designed to compare samples that are already close: a predicted vs ground truth depth map, or two clips of similar motion. We therefore use d only to find each sample's nearest neighbors and to weight those edges. The normalized Laplacian of the resulting graph is positive semidefinite for any symmetric non-negative weights, and its heat semigroup gives an explicit Euclidean embedding, the diffusion embedding. Any standard distri-

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.

butional statistic computed on this embedding is then valid, and energy distance, which needs no bandwidth, is our default.

A second requirement comes from how evaluation scores are used. A benchmark ranks many generators, so all of them must be scored in the same diffusion embedding. Methods that build a graph or an embedding on the combined real and generated samples give every comparison its own geometry, and adding a generator can change the scores of the others. We therefore fit the graph once, on real data alone, and place every evaluation and generated sample on it by the Nyström extension, so each generator's score is a function of only that generator's samples and the fixed real data.

We call the resulting score the Diffusion Energy Divergence (DED), but we clarify that "diffusion" refers to heat diffusion on the graph, not to diffusion models. Our contributions are as follows:

- **Failure of negative type.** We show that the distances used in many fields, from Chamfer distance to dynamic time warping, fail negative type on real data, and for example, energy distance on AbsRel ranks two depth degradations in reverse order of severity.
- **Meaningful local distances.** DED accepts a far larger class of distances than energy distance and distance-induced MMD, and depends on a distance only through each sample's κ nearest neighbors and its values on those pairs (Proposition 1).
- **Fixed geometry.** Unlike graph-based comparisons built on the pooled samples, we fit the graph once on real data and insert every other sample with the Nyström extension, so every generator is scored in the same diffusion embedding. This also gives a permutation test that combines the divergence with the fraction of samples placed off the real data (Proposition 3).
- **Experimental results.** We demonstrate DED on depth maps and video, and on audio and a ViT feature space in the appendix. On VBench, a clip's first frame repeated beats the real clip it came from on subject consistency and flickering, and DED detects in 80% of the generators a mode collapse that leaves every VBench mean unchanged in expectation.

2 Related Work

Feature-based evaluation. FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) compare Inception activations, CMMD (Jayasumana et al., 2024) compares CLIP embeddings under a Gaussian kernel, and FVD (Unterthiner et al., 2019) applies the Fréchet distance to features of a supervised action-recognition network. Each inherits the geometry of its encoder. FVD computed on I3D features responds mainly to per-frame appearance, and the same distance on self-supervised VideoMAE features reduces this bias (Ge et al., 2024). JEDi (Luo et al., 2025) replaces both the features and the Gaussian assumption. These approaches change the features or the statistic. We keep the choice of distance, and change how a divergence is built from it.

Video evaluation. STREAM (Kim et al., 2024), FVMD (Liu et al., 2024) and TRAJAN (Allen et al., 2025) separate motion from appearance with purpose-built representations (the temporal spectrum of frame embeddings, histograms of point-track velocities, and a point-track autoencoder), but none of them is a divergence with a permutation test, and DED can take a distance built on any of them. VBench (Huang et al., 2023) averages per-video scores without reference to real data. Its authors note that a static video scores well on its temporal dimensions and add dynamic degree as a counterweight, and Section 6 and Appendix L measure how far this problem extends.

Distributional comparison under a given distance. Generalized energy distance (Székely & Rizzo, 2013) and distance-induced MMD (Sejdinovic et al., 2013) require negative type (Lyons, 2013; Schoenberg, 1938; Berg et al., 1984), and the proof that Sinkhorn divergences are positive assumes it (Feydy et al., 2018). Graph-based two-sample tests (Friedman & Rafsky, 1979; Chen & Friedman, 2013) accept any distance, and classifier tests (Lopez-Paz & Oquab, 2016) any features, but both are refitted on the pooled data of each comparison, so they give a decision for one pair of datasets rather than a score on a shared scale. The same holds for 1-NNA, which point-cloud work reports alongside MMD-CD and COV-CD (Achlioptas et al., 2018; Yang et al., 2019). These are a nearest-neighbor average and a coverage count on Chamfer distance, and neither is a divergence.

Graph diffusion for comparing distributions. Laplacian eigenmaps, diffusion maps and diffusion kernels are standard (Belkin & Niyogi, 2003; Coifman & Lafon, 2006a; Kondor & Lafferty, 2002), and several works compare distributions with them. Diffusion EMD (Tong et al., 2021) and Geodesic Sinkhorn (Huguet et al., 2022) build their graph on the pooled datasets, Graph Fourier

MMD (Leone et al., 2023) needs distributions supported on the vertices of the graph, IMD (Tsitsulin et al., 2019) builds a separate graph for each dataset and so cannot say where two datasets lie relative to each other, and the reference-set MMD of Cheng et al. (2017), the closest to our setting, needs Euclidean inputs. Energy distance on the diffusion embedding is itself equivalent to an MMD, with the distance-induced kernel of a truncated diffusion distance (Sejdinovic et al., 2013). What is new is how the components are used: an arbitrary distance enters only through nearest neighbors, the graph is fitted once on real data, and every other sample is placed on it by the Nyström extension (Bengio et al., 2003; Coifman & Lafon, 2006b), with the samples off the graph tested jointly with the divergence.

3 Background

3.1 Energy Distance and MMD

Let $X, X' \sim P$ and $Y, Y' \sim Q$ be independent, with finite first moments. The energy distance (Székely & Rizzo, 2013) between P and Q is

$$D^2(P,Q) = 2\mathbb{E}\|X-Y\|_2 - \mathbb{E}\|X-X'\|_2 - \mathbb{E}\|Y-Y'\|_2 \tag{1}$$

where $\|\cdot\|_2$ is the Euclidean norm. It has two properties that matter here: $D^2(P,Q) \geq 0$, and $D^2(P,Q) = 0$ only if $P = Q$, which follows from the strong negative type of Euclidean space (Section 3.2). Hence D^2 is a divergence between distributions and its square root D is a metric (Székely & Rizzo, 2013).

MMD (Gretton et al., 2012) compares two probability measures through a positive-definite kernel k: with mean embeddings $\mu_P = \mathbb{E}_{X\sim P}[k(X,\cdot)]$ in the Reproducing Kernel Hilbert Space $\mathcal{H}_k$, $\mathrm{MMD}_k(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$. It is a metric on probability measures exactly when $P \mapsto \mu_P$ is injective, and kernels with this property are called *characteristic* (Sriperumbudur et al., 2010). CMMD (Jayasumana et al., 2024), an MMD on CLIP embeddings, uses a Gaussian kernel, whose bandwidth is one more hyperparameter to tune. Under the distance kernel $k(x,y) = \frac{1}{2}(\|x\|_2 + \|y\|_2 - \|x-y\|_2)$, $\mathrm{MMD}_k(P,Q)^2 = \frac{1}{2}D^2(P,Q)$, so energy distance is, up to a factor of two, the squared MMD of a kernel with no bandwidth (Sejdinovic et al., 2013). Appendix J gives the details and the unbiased estimator.

Replacing the Euclidean norm by a distance d gives the generalized energy distance

$$D_d^2(P,Q) = 2\mathbb{E}[d(X,Y)] - \mathbb{E}[d(X,X')] - \mathbb{E}[d(Y,Y')]. \tag{2}$$

A modality on which Euclidean distance is meaningless needs this generalized energy distance, but it is a valid divergence (and in fact guaranteed to even be non-negative) only when d is of negative type.

3.2 Negative Type

Definition 1. *A distance d on a set M is of negative type if for any integer $n \geq 2$, any collection of points $x_1, \ldots, x_n \in M$ and any coefficients $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$, the following inequality holds:*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j d(x_i, x_j) \leq 0 \tag{3}$$

It is of strong negative type if, in addition, $D_d^2(P,Q) = 0$ implies $P = Q$ for all probability measures P, Q on M with finite first moment.

Negative type constrains every finite configuration of points, so a single distorted long-range distance can make it fail. Table 15 measures it for many commonly used distances, and all of them fail on real data except SILog and exact optimal transport between token sets, which pass up to rounding error. Appendix M describes the test and the connection to positive-definite kernels through Schoenberg's theorem.

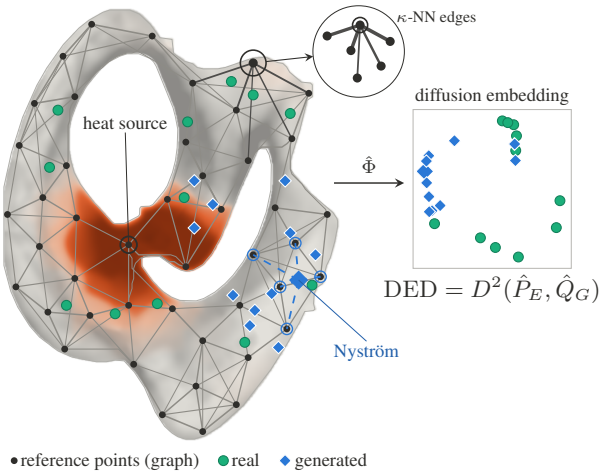

Figure 1: **DED on a surface with two holes.** Reference points (**black**) are joined to their κ nearest neighbors under the distance d (inset). Heat e^{-tL} released at the heat source (orange) spreads along these edges, around the holes rather than across them. A new sample (large diamond) is placed from its κ nearest reference points (circled) by the Nyström extension, which leaves the graph unchanged. The generated samples (**blue**) cover only the lower right of the surface, so $\hat{\Phi}$ separates them from the real samples (**green**) in the diffusion embedding, where DED is the energy distance between them.

4 The Diffusion Energy Divergence

Given samples $x_1, \dots, x_{n_P} \sim P$ of real data and $y_1, \dots, y_{n_Q} \sim Q$ from a generative model, evaluating the model means computing a two-sample statistic that measures how far Q is from P. Samples are usually high-dimensional, so standard methods usually first embed them with a pretrained encoder. FID (Heusel et al., 2017) fits a Gaussian to each set of embeddings and compares the two fits with the 2-Wasserstein distance, while CMMD (Jayasumana et al., 2024) drops the Gaussian assumption by using MMD on CLIP embeddings. Here, we introduce *Diffusion Energy Divergence* (DED) as a stronger and more versatile alternative.

Heat Kernels on a Manifold. Let M be a compact Riemannian manifold with Laplace-Beltrami operator $\Delta = -\mathrm{div}\,\mathrm{grad}$. The heat equation $\partial_t u = -\Delta u$ has solution operator $e^{-t\Delta}$, whose integral kernel $p_t(x, y)$, the **heat kernel**, is the density at y after time t of heat released at x. Using the orthonormal eigenfunctions φ_k of Δ, with eigenvalues $0 = \lambda_0 \le \lambda_1 \le \cdots$, it can be expanded as $p_t(x, y) = \sum_k e^{-t\lambda_k} \varphi_k(x)\varphi_k(y)$, a sum of positive multiples of outer products, so with $\Phi_t(x) = (e^{-t\lambda_k/2}\varphi_k(x))_{k\ge 0}$, $p_t(x, y) = \langle \Phi_t(x), \Phi_t(y)\rangle$. The heat kernel is therefore an inner product between explicit coordinates, whatever the geometry. Small t sees only local geometry, larger t progressively coarser structure, and as $t \to \infty$ only the constant eigenfunction survives. Appendix B gives the details, a more thorough explanation of these concepts, and the small-t limit.

Heat Kernels on a Sample Graph. In practice we have samples, not a manifold. Given samples $x_1, \dots, x_n$ and a symmetric weight matrix $W \ge 0$, we construct a weighted graph upon which we form the normalized graph Laplacian $L = I - D^{-1/2} W D^{-1/2}$ (Von Luxburg, 2007; Ng et al., 2001), where $D = \mathrm{diag}(\deg_i)$ and $\deg_i = \sum_j W_{ij}$. Like Δ, it is positive semidefinite:

$$\mathbf{a}^\top L \mathbf{a} = \sum_{i<j} W_{ij} \left(\frac{a_i}{\sqrt{\deg_i}} - \frac{a_j}{\sqrt{\deg_j}} \right)^2 \ge 0 \tag{4}$$

for every $\mathbf{a} \in \mathbb{R}^n$ and every symmetric $W \ge 0$ with positive degrees. Writing $L = \sum_k \lambda_k u_k u_k^\top$, the heat semigroup $K_t = e^{-tL} = \sum_k e^{-t\lambda_k} u_k u_k^\top$ is positive definite (Kondor & Lafferty, 2002), and its eigenvectors give an explicit embedding.

We use the diffusion-map normalization of this embedding (Coifman & Lafon, 2006a). Let $T = D^{-1}W$ be the random walk on the graph, with stationary distribution $\pi_i = \deg_i / \sum_j \deg_j$. Its right eigenvectors are $\psi_k = u_k / u_0$, with eigenvalues $\mu_k = 1 - \lambda_k$, normalized so that $\sum_i \pi_i \psi_k(i)^2 = 1$ (ψ_0 is constant and is dropped). The diffusion embedding of a reference point is

$$\Phi_t(x_i) = \left(\mu_k \, e^{-t\lambda_k/2} \, \psi_k(i) \right)_{k=1}^{m}, \tag{5}$$

over the m modes with the smallest λ_k. The distance between two diffusion embeddings is the diffusion distance of Coifman & Lafon (2006a) for the operator $A_t = T e^{-t(I-T)/2}$: $\|\Phi_t(x_i) - \Phi_t(x_j)\|_2$

is the (truncated) $L^2(1/\pi)$ distance (where $\|f\|^2_{L^2(1/\pi)} = \sum_l f(l)^2/\pi_l$) between the distributions reached by taking one step of the walk from x_i or from x_j and then diffusing for time $t/2$. The factor μ_k lets a new sample z be placed by the same rule, giving $\hat{\Phi}_t(z)$. We concatenate 3 different diffusion times into $\hat{\Phi}(z) \in \mathbb{R}^{3m}$ (Section 5). We call Φ_t the **diffusion embedding** and the energy distance between the distributions of real and generated samples in it the **Diffusion Energy Divergence (DED)**. Like energy distance D^2, DED is a divergence rather than a metric: it is non-negative and zero only when the two distributions of diffusion embeddings coincide (Remark 1). Figure 1 shows DED on a toy surface.

Graph Construction from a Distance. Fix a set of real reference samples $R = \{x_1, \ldots, x_n\}$. Let $\mathcal{N}_\kappa(i)$ be the indices of the κ nearest neighbors of x_i in R under d, and $\sigma_i = \max_{j \in \mathcal{N}_\kappa(i)} d(x_i, x_j)$ the distance to the κ-th of them. The pair (i, j) is a *retained edge* if $j \in \mathcal{N}_\kappa(i)$ or $i \in \mathcal{N}_\kappa(j)$, and N denotes the set of retained edges. Following self-tuning spectral clustering (Zelnik-Manor & Perona, 2004), the weights are $\tilde{W}_{ij} = \exp(-d(x_i, x_j)^2/(\sigma_i\sigma_j))$ on retained edges and 0 elsewhere. We then apply the density normalization of Coifman & Lafon (2006a): $q_i = \sum_j \tilde{W}_{ij}$ and $W_{ij} = \tilde{W}_{ij}(q_i q_j)^{-\alpha}$, with $\alpha = 1$. With a fixed bandwidth, $\alpha = 1$ removes the sampling density from the limiting operator. However, with bandwidths that adapt to the local sampling density, as here, the limit is a density-weighted operator instead (Ting et al., 2011; Berry & Harlim, 2014). Nothing below relies on the limit, and $\alpha \in \{0, 0.5\}$ gives the same rank correlations as $\alpha = 1$ on every depth degradation (Appendix I).

4.1 Locality and Validity

DED defines a new geometry from the neighborhood structure of d, in which two samples that are far apart under d can be close if many short paths connect them. The first result states how little of d this geometry uses.

Proposition 1 (Locality). *K_t and the diffusion embeddings of the reference points are determined by the neighbor lists $\{\mathcal{N}_\kappa(i)\}_{i=1}^n$ together with the restriction $d|_N$ (the embeddings up to the choice of eigenbasis, which leaves DED unchanged when $\lambda_m < \lambda_{m+1}$). The diffusion embedding $\hat{\Phi}_t(z)$ of a new sample z (Equation 6) is determined in addition by $\mathcal{N}_\kappa(z)$ and $d|_{N_z}$, where $N_z = \{z\} \times \mathcal{N}_\kappa(z)$. Hence if d' induces the same neighbor lists and agrees with d on N and on every N_z, then $K'_t = K_t$ and $\hat{\Phi}'_t = \hat{\Phi}_t$ for every $t > 0$.*

Increasing the distance between two samples that are not neighbors cannot change any neighbor list, so it cannot change the score. However, if the distance between two non-neighbors falls below the distance from one of them to its current κ-th neighbor, the pair becomes an edge and the embedding changes. The proposition therefore says only that the score ignores the values of d between non-neighbors.

Proposition 2 (Negative type is not local). *Let $a, b, c \in R$ be distinct, with $(a, b) \notin N$ and $d(a, b) > 0$. Define d' to equal d except that $d'(a, b) = d'(b, a) = F\, d(a, b)$ with $F > \max\{1,\, 2\,(d(a, c) + d(b, c))/d(a, b)\}$. Then d' has the same neighbor lists as d and agrees with it on N, and for $P = \frac{1}{2}(\delta_a + \delta_b)$ and $Q = \delta_c$ the generalized energy distance satisfies $D^2_{d'}(P, Q) = d(a, c) + d(b, c) - \frac{1}{2}F d(a, b) < 0$.*

Hence, whenever the graph is not complete, every distance has a version with the same neighbor lists and the same values on neighboring pairs on which energy distance goes negative. By Proposition 1, the diffusion embedding is identical for the two versions. Appendix E shows that a distance that is exactly Euclidean on every retained edge drives the generalized energy distance negative, while DED, which by Proposition 1 equals DED on the Euclidean distance, rises with severity. Both proofs are in Appendix F.

Remark 1 (Validity). (1) *$L \succeq 0$ for any symmetric non-negative W with positive degrees, so K_t has eigenvalues in $(0, 1]$ for every $t > 0$.* (2) *The diffusion embedding is an explicit Euclidean map, and its Gram matrix $\Phi_t\Phi_t^\top = \sum_{k=1}^m \mu_k^2 e^{-t\lambda_k}\psi_k\psi_k^\top$ is positive semidefinite.* (3) *For any bounded map $\hat{\Phi}$ into a Euclidean space, such as the diffusion embedding of Section 5, and with $\hat{\Phi}_\# P$ the distribution of $\hat{\Phi}(X)$ for $X \sim P$, $D^2(\hat{\Phi}_\# P, \hat{\Phi}_\# Q) \geq 0$, with equality if and only if $\hat{\Phi}_\# P = \hat{\Phi}_\# Q$, because Euclidean space is of strong negative type (Lyons, 2013).*

A divergence between which distributions? Remark 1 concerns the distributions of diffusion embeddings. DED is a divergence between P and Q themselves only when $\hat{\Phi}$ is injective on the union of their supports, which depends on two things. The first is the number of modes. With all $n-1$ nontrivial modes retained, the vectors of all n reference points in Equation 6 are affinely independent, so $\hat{\Phi}_t(z)$ determines the neighbor weights $p(z,\cdot)$ exactly. Truncating to m modes, which we do for efficiency (otherwise the dimensionality of each embedding equals the number of reference points), can make samples with different weights collide. The second is the out-of-sample rule: the weights $p(z,\cdot)$ depend only on which reference points are z's neighbors and on the relative weights of those edges (Equation 6), so a sample and a much more distant one with the same neighbor pattern collide whatever the number of modes. DED is therefore a divergence between the distributions of diffusion embeddings, and Section 5.1 detects collisions of the second kind with a second statistic tested jointly with the divergence.

4.2 Coordinate Statistics

Because the diffusion embedding is Euclidean, any two-sample statistic can be computed on it. The MMD of the kernel $\Phi_t\Phi_t^\top$ is a logical first choice, but this is the linear kernel on the diffusion embedding, and its squared MMD, $\|\mathbb{E}_P[\hat{\Phi}_t(X)] - \mathbb{E}_Q[\hat{\Phi}_t(Y)]\|_2^2$, compares only the m coordinate means. (With every mode retained, every $\mu_k \neq 0$ and every sample a reference point, these means would determine the distribution.) Energy distance on the same embedding is, up to a factor of two, the squared MMD of the distance kernel on $\mathbb{R}^{3m}$, which is characteristic, so it compares the entire distributions of $\hat{\Phi}(X)$ and $\hat{\Phi}(Y)$ whatever the map $\hat{\Phi}$ is, and needs no bandwidth. FID, KID or a Gaussian-kernel MMD can be computed on the same embedding for comparison.

5 Applying DED

The method needs a distance d that is non-negative, symmetric, zero on the diagonal and meaningful for nearby samples, a set of real samples, and generators $Q_1, \ldots, Q_C$ to evaluate. What counts as meaningful depends on the modality and the task, and d need not satisfy the triangle inequality or be of negative type. The real data are split into two disjoint sets: a **reference set** R, $|R| = n$, which builds the graph and is never scored, and an **evaluation set** E, $|E| = n_E$, which generated samples are compared against. If the two sets overlapped, the real data would lie directly on the vertices of the graph but the generated data would not (even if it matches the distribution perfectly), and the resulting gap would be attributed to the generator.

Fitting the Reference Graph. Everything derived from R (neighbor lists, bandwidths, density factors, spectrum) is computed once and never changed by the samples being scored. We compute the distances $d(x_i, x_j)$ within R, then W, L and its spectrum. We keep the $m = 64$ modes with the smallest nonzero eigenvalues. The diffusion times come from the retained spectrum: $t_1 = 1/\lambda_{\max}$ and $t_3 = 1/\lambda_{\min}$, with $\lambda_{\max}$ and $\lambda_{\min}$ the largest and smallest retained eigenvalues, and $t_2 = \sqrt{t_1 t_3}$. Algorithm 1 and Appendix G.1 give the full procedure, including what to do when the graph is disconnected.

Out-of-Sample Placement. The evaluation and generated samples are placed on the fixed graph without changing it, so that every generator is scored in the same diffusion embedding. A new sample z is attached through the indices $\mathcal{N}_\kappa(z)$ of its κ nearest reference points, with the same formulas as a reference point: $\sigma_z = \max_{j\in\mathcal{N}_\kappa(z)} d(z, x_j)$, $\tilde{W}_{zj} = \exp(-d(z,x_j)^2/(\sigma_z\sigma_j))$ on $\mathcal{N}_\kappa(z)$, $q_z = \sum_j \tilde{W}_{zj}$, $W_{zj} = \tilde{W}_{zj}(q_z q_j)^{-\alpha}$ and $\deg_z = \sum_j W_{zj}$, with every reference quantity held fixed. With $p(z,j) = W_{zj}/\deg_z$, the Nyström extension of the random-walk eigenvectors (Bengio et al., 2003; Coifman & Lafon, 2006b) is $\psi_k(z) = \mu_k^{-1}\sum_j p(z,j)\,\psi_k(j)$. The factor μ_k in Equation 5 cancels the division, so

$$\hat{\Phi}_t(z) = \sum_{j\in\mathcal{N}_\kappa(z)} p(z,j)\left(e^{-t\lambda_k/2}\,\psi_k(j)\right)_{k=1}^{m}, \tag{6}$$

that is, z is placed at a convex combination of its neighbors' vectors $v_j = (e^{-t\lambda_k/2}\psi_k(j))_{k=1}^m$. Applied to a reference point with its fitted row $p(i,\cdot)$, this returns $\Phi_t(x_i)$, and is Lipschitz in the neighbor weights with no amplification (Proposition 4 in Appendix G.2). The standard Nyström rule

divides by μ_k instead, which amplifies modes with λ_k near 1 and needs a spectral cut. Restricting the sum to $\mathcal{N}_\kappa(z)$ keeps the score local (Proposition 1). Appendix G.2 gives the derivation.

The Estimator. Every sample z in E or in a generated set is mapped to $\hat{\Phi}(z) = (\hat{\Phi}_{t_1}(z), \hat{\Phi}_{t_2}(z), \hat{\Phi}_{t_3}(z)) \in \mathbb{R}^{3m}$, which is fixed by the reference graph before any of these samples are seen. For $E = \{e_1, \dots, e_{n_E}\}$ and a generated set $G = \{g_1, \dots, g_{n_G}\}$ the empirical measures of their diffusion embeddings are

$$\hat{P}_E = \frac{1}{n_E} \sum_{i=1}^{n_E} \delta_{\hat{\Phi}(e_i)}, \qquad \hat{Q}_G = \frac{1}{n_G} \sum_{j=1}^{n_G} \delta_{\hat{\Phi}(g_j)}, \tag{7}$$

that is, $\hat{P}_E(A) = \frac{1}{n_E} \#\{i : \hat{\Phi}(e_i) \in A\}$ for $A \subseteq \mathbb{R}^{3m}$. The population quantity is $\mathrm{DED}(P, Q) = D^2(\hat{\Phi}_\# P, \hat{\Phi}_\# Q)$. Write $\bar{d}_{EG}$ for the mean of $\|\hat{\Phi}(e_i) - \hat{\Phi}(g_j)\|_2$ over all pairs, and $\bar{d}_{EE}$, $\bar{d}_{GG}$ for the means over distinct pairs within E and within G. Replacing the two pushforwards by $\hat{P}_E$ and $\hat{Q}_G$ gives the plug-in estimate $D^2(\hat{P}_E, \hat{Q}_G)$, the energy distance between the two empirical measures. It is non-negative but biased upward, because its within-sample averages include the zero diagonal terms. Removing them gives the unbiased U-statistic

$$\widehat{\mathrm{DED}} = 2\,\bar{d}_{EG} - \bar{d}_{EE} - \bar{d}_{GG}, \tag{8}$$

Because $\hat{\Phi}$ is fitted on R alone, this is an ordinary two-sample U-statistic on Euclidean data, and a permutation test on the pooled diffusion embeddings is exact under $P = Q$. We report the z-score $(\widehat{\mathrm{DED}} - \text{null mean})/\text{null s.d.}$ over 200 permutations of $E \cup G$ as an effect size, and decide significance with the combined test below.

5.1 Samples Off the Reference Support

The placement depends only on the relative weights of a sample's neighbors, so a sample far from every reference point is placed where a nearby sample with the same neighbor pattern would be, and the divergence cannot see how far away it is. The log affinity mass $h(z) = \log \sum_{j \in \mathcal{N}_\kappa(z)} \tilde{W}_{zj}$ does see it, because it decreases without bound as z moves away from the reference. We call z **off-support** when $h(z)$ falls below the 1% quantile τ of the reference points' own values (each computed against the other $n - 1$), and test the off-support rate jointly with the divergence.

Proposition 3 (Combined test)**.** *On random re-splits of $E \cup G$ at sizes (n_E, n_G), compute the U-statistic of Equation 8 and the difference between the off-support rates of G and E, turn each into a permutation p-value, p_D and p_O, and let p be the permutation p-value of $\min(p_D, p_O)$. If $P = Q$ then $\Pr(p \le a) \le a$ for every a, whatever d is. With B permutations, the test is consistent at every level $a \ge 2/(B+1)$ if $\hat{\Phi}_\# P \ne \hat{\Phi}_\# Q$ or if Q puts more mass on $\{h < \tau\}$ than P does.*

The size holds because $\hat{\Phi}$ and τ are fitted on R alone, so under $P = Q$ the pooled pairs $(\hat{\Phi}(z), \mathbf{1}[h(z) < \tau])$ are exchangeable. Consistency follows from $p \le 2\min(p_D, p_O)$ (Appendix F). Since $h(z)$ and τ depend on d only through $d|_{N_z}$ and $d|_N$, the argument of Proposition 1 covers the second statistic too.

6 Experiments

Setup. We ask three questions: whether DED ranks controlled degradations of real data in order of severity when the distance fails negative type, whether it responds correctly to the settings of a real generator, and what it detects on a standard video benchmark that a mean of per-video scores cannot. For the first we degrade held-out real samples at increasing severity: 32×32 depth patches from NYU Depth V2 (Silberman et al., 2012) under the symmetrized AbsRel ($|R| = 300$, $|E| = 150$ and degraded pools of 150), and clips from RealCam-Vid (Zheng et al., 2025b), processed at 7 frames per second, under dynamic time warping between optical-flow profiles. For the second we generate clips with Stable Video Diffusion (Blattmann et al., 2023) from the first frames of RealCam-Vid clips and vary its sampler steps and motion strength. For the third we use the released samples of 20 generators from VBench (Huang et al., 2023), with MSR-VTT clips (Xu et al., 2016) as the reference. In every experiment R and E are disjoint sets of real samples, and the graph is fitted once.

Table 1: **Video: damaging real clips in controlled ways.** Each row applies one kind of damage to real clips at five increasing severities. Color jitter changes only frame brightness, shuffle and reorder change only frame order, and freeze, drop-hold and blur change both. The top part shows how well each score follows the severity, as a Spearman correlation ρ. A value in parentheses means the score never moved beyond its noise level ($|z| < 2$ at every severity), which we measure by randomly re-splitting the clips, so its ranking carries no information. The bottom part shows the largest move of each score, in standard deviations of that noise. The rows to look at are shuffle and reorder: FVD (with S3D features) cannot tell them apart from noise, while DED detects both clearly. Under color jitter the roles reverse, and DED stays flat, as desired.

	FVD	DED
appearance		
jitter	+1.00	(−0.40)
temporal (pure)		
shuffle	(+1.00)	**+1.00**
reorder	(+1.00)	**+1.00**
mixed		
freeze	+1.00	+1.00
drop-hold	+1.00	+1.00
blur	+1.00	+1.00
max $\|z\|$ jitter	3.5	**0.4**
max $\|z\|$ shuffle	1.7	**23.8**
max $\|z\|$ reorder	1.0	**7.6**
max $\|z\|$ mixed	25.3	**153.1**

Table 2: **Depth maps: corrupting NYU depth patches.** Each column is one kind of damage applied at seven severities. Collapse replaces patches with copies of a few patches, corrupt blurs and adds noise, and scale multiplies each map by a random factor. Each entry is the Spearman correlation ρ between a score and the severity. The SILog baselines ignore overall scale, so they stay exactly constant there. Energy distance computed directly on AbsRel does see scale, but it is not a valid divergence, and on two of the three columns it gets the order backwards (-0.89), judging the most damaged pool as the closest to the real data. DED built on the same AbsRel distance ranks all three correctly.

	collapse	corrupt	scale
Baselines			
energy (SILog)	+1.00	+1.00	const
Fréchet (SILog)	+1.00	+1.00	const
Ablation			
energy (AbsRel)	+1.00	$\mathbf{-0.89}$	$\mathbf{-0.89}$
Ours			
DED	**+1.00**	**+0.96**	**+1.00**
off-support (max)	0.01	1.00	0.25

Table 3: **Video: Stable Video Diffusion clips.** We vary two settings of the generator, the number of sampler steps and the motion strength. Each entry is the Spearman correlation ρ between a setting and the score with its sign flipped. Motion strength does not make video better or worse, but it shows how each score treats motion. Both FVD variants prefer the low-motion output of the lowest setting, while DED judges that output furthest from real video. DED here uses the optical-flow distance of Table 1.

	sampler steps	motion
FVD (S3D)	**+0.89**	−1.00
FVD (VideoMAE)	**+0.89**	−0.80
DED	+0.60	**+0.80**

The baselines are the standard scores of each field: energy distance and the Fréchet distance on SILog for depth, FVD on S3D and VideoMAE (Ge et al., 2024) features for video, and VBench's temporal dimensions. For depth we also report energy distance computed directly on AbsRel, which is the construction DED replaces. The tables report the Spearman correlation ρ between a score and the degradation severity or generator setting, and where noted z, the number of standard deviations a score lies from its distribution when the two sets being compared come from the same source. The appendix adds a ViT feature space, two-dimensional examples, the video settings and audio (Appendices C, E, H and K). Unless stated otherwise, $\kappa = 8$ with the self-tuning bandwidth, $\alpha = 1$, at most $m = 64$ modes, and three diffusion times taken from the spectrum of L.

Depth. Generative models increasingly produce depth (Bai et al., 2026; Weijler et al., 2026), and a generated depth map has no paired ground truth, so per-sample errors such as AbsRel and SILog cannot score a generator on their own. SILog, the Euclidean distance between mean-centered log depths, is of negative type, and energy distance applies to it directly, but it is scale-invariant, so it cannot see a difference in scale. AbsRel measures depth at its true scale but is not of negative type, so energy distance on it carries no guarantee. Since DED needs a symmetric distance, we use $d(A, B) = \frac{1}{2}(\mathrm{AbsRel}(A, B) + \mathrm{AbsRel}(B, A))$. We apply three degradations, each at seven

severities q from 0 to 1: *collapse* replaces a fraction q of the data with copies of a few real samples, *corrupt* applies local smoothing and multiplicative noise that grow with q, and *scale* multiplies each sample by a random factor whose spread grows with q. Energy distance on AbsRel ranks both the corruption and the scale degradation in reverse order of severity ($\rho = -0.89$, Table 2). DED on the same distance positively correlates with the severity of all three degradations ($\rho \geq 0.96$), including scale, which SILog cannot see. At the lowest nonzero corruption severity, 78% of the corrupted patches have already left the reference support. Each is placed where a clean patch with the same neighbors would be, so the energy test alone finds no difference ($p = 0.99$), while the combined test of Proposition 3, which also counts off-support samples, does ($p = 0.01$).

Video. A video generator can render plausible frames with the wrong motion, and scores computed on pooled encoder features, such as FVD, cannot separate errors of motion from errors of appearance. We use dynamic time warping (Berndt & Clifford, 1994) between optical-flow profiles, which aligns two clips in time before comparing their motion and is not of negative type. We apply one appearance corruption (color jitter), two purely temporal ones (shuffle and local reorder) and three that affect both (freeze, drop-hold and temporal blur), each at five severities (Table 1). Under jitter DED stays within its noise level at every severity ($|z| \leq 0.4$), and it rises with every temporal and mixed corruption. FVD rises with jitter and with the three mixed corruptions, but under shuffle and reorder it never leaves its noise level ($|z| \leq 1.7$ and 1.0, against 23.8 and 7.6 for DED). On Stable Video Diffusion (Table 3) both scores improve with the number of sampler steps, FVD more closely. DED separates one or two steps from four or more ($z \approx 61$ against 8 to 10) but does not order the settings from four steps up. On motion strength both FVD variants prefer the low-motion output of the lowest setting, while DED places it furthest from real video. Appendix K finds the same division in audio, with timbre in place of appearance.

A standard video benchmark. VBench (Huang et al., 2023) scores many aspects of generated video, and its temporal dimensions are per-video statistics that never see real footage. A mean of per-video scores is not a divergence, and these dimensions pull in opposite directions: subject consistency and flickering are best for a still image, dynamic degree for constant change. Scored with VBench's official code, a clip's first frame repeated beats the real clip it came from on subject consistency (99.7% of 300 clips) and on temporal flickering (100%). Fresh noise on every frame is scored as dynamic on every clip and beats real footage on subject consistency for 89.7% of the clips. On the 20 generators' own clips, replacing each prompt's samples by copies of one of them (mode collapse) leaves every VBench mean unchanged in expectation, and none of the three VBench temporal scores we recompute per clip detects it in more than 20% of the models. Duplicating every other frame (stutter) improves the flickering score in all 20. DED compares each manipulated pool with the same generator's clean pool, so the reference does not need to cover the generator. Under DTW on the drift of DINOv2 (Oquab et al., 2023) features across frames, it detects that collapse in 80% of the models, and time reversal (the clip played backwards), ping-pong playback (the first half played forward and then back) and stutter in all of them. Appendix L gives the setup, the leaderboard correlations and DED's ranking of the 20 models.

7 Conclusion

Energy distance and distance-induced MMD require the distance they are built on to be of negative type, and many of the distances that depth, point clouds, motion, audio and token sets are compared with fail that condition on real data. We use such a distance only to connect each sample to its nearest neighbors. The heat semigroup of the resulting graph gives a Euclidean diffusion embedding in which energy distance is valid, and this embedding depends on the distance only through the neighbor lists and the values on neighboring pairs. A permutation test on the embedding and the off-support rate has exact size for any distance. On depth this corrects two inverted rankings and recovers the scale errors missed by SILog. On video, it separates motion from appearance. On VBench, it detects a mode collapse that leaves a mean of per-video scores unchanged and a stutter that such a mean rewards.

Limitations and future work. Two limitations follow from the construction itself. If d is not reliable even between neighbors, the diffusion embedding is still valid but no longer informative. And a ranking is relative to its reference set, which has to cover the generated samples and should be stated along with the scores. Future work will seek to ensure that even off-support samples can be properly embedded, to combine the two needed scores into one.

AI USE STATEMENT

In this work, we used generative AI tools to assist with parts of the coding, to polish text and to double-check our proofs for correctness. We have not used generative AI tools to write large parts of the paper, to prepare citations or to produce the research ideas themselves. We have reviewed all AI-assisted work. This includes manually checking any written code, as well as checking polished text to ensure that it remains accurate to our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Algorithm 1 and Appendix G.1 give the full procedure, and Section 6 and the appendices give the settings of every experiment. We plan to release the code for this paper at the end of the review period.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pp. 40–49. PMLR, 2018.

Kelsey Allen, Carl Doersch, Guangyao Zhou, Mohammed Suhail, Danny Driess, Ignacio Rocco, Yulia Rubanova, Thomas Kipf, Mehdi S. M. Sajjadi, Kevin Murphy, Joao Carreira, and Sjoerd van Steenkiste. Direct motion models for assessing generated videos. *ArXiv*, abs/2505.00209, 2025.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *Advances in Neural Information Processing Systems*, 38:57602–57622, 2026.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel pca. *Neural computation*, 16(10):2197–2219, 2004.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. Harmonic analysis on semigroups: Theory of positive definite and related functions. 1984.

Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pp. 359–370, 1994.

Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *arXiv: Spectral Theory*, 2014.

Tyrus Berry and Tim Sauer. Local kernels and the geometric structure of data. *arXiv: Classical Analysis and ODEs*, 2014.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Andrew Bond, Ilkin Umut Melanlioglu, Erkut Erdem, and Aykut Erdem. Beyond gaussian bottlenecks: Topologically aligned encoding of vision-transformer feature spaces. *arXiv preprint arXiv:2604.28122*, 2026a.

Andrew Bond, Ege Erdem Özlü, Tuna Çimen, Ilkin Umut Melanlioglu, Tolga Birdal, Erkut Erdem, and Aykut Erdem. Task-induced Riemannian metrics for vision transformer feature spaces. *arXiv preprint arXiv:2609.27988*, 2026b.

Jeff Calder and Nicolas Garcia Trillos. Improved spectral convergence rates for graph laplacians on epsilon-graphs and k-nn graphs, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Hao Chen and Jerome H. Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112:397 – 409, 2013.

Xiuyuan Cheng, Alexander Cloninger, and Ronald R. Coifman. Two-sample statistics based on anisotropic kernels. *Information and inference : a journal of the IMA*, 9 3:677–719, 2017.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006a.

Ronald R. Coifman and Stéphane Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21: 31–52, 2006b.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist.*, 7(1):697–717, 1979.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288. IEEE, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Xuan He, Dongfu Jiang, Ge Zhang, Max W.F. Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi "Richard" Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(6), 2007.

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. *Advances in neural information processing systems*, 30, 2017.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qin Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2023.

Guillaume Huguet, Alexander Tong, María Ramos Zapatero, Christopher J. Tape, Guy Wolf, and Smita Krishnaswamy. Geodesic sinkhorn for fast and accurate optimal transport on manifolds. *ArXiv*, 2022.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Pum Jun Kim, Seojun Kim, and Jaejun Yoo. Stream: Spatio-temporal evaluation and analysis metric for video generative models. *ArXiv*, abs/2403.09669, 2024.

Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, volume 2002, pp. 315–322, 2002.

Samuel Leone, Aarthi Venkat, Guillaume Huguet, Alexander Tong, Guy Wolf, and Smita Krishnaswamy. Graph fourier mmd for signals on graphs, 2023.

Jia-He Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. *ArXiv*, abs/2407.16124, 2024.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv: Machine Learning*, 2016.

Ge Ya Luo, Gian Mario Favero, Zhi-Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Joseph Pal. Beyond fvd: An enhanced evaluation metrics for video generation distribution quality. In *International Conference on Learning Representations*, 2025.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, pp. 3284–3305, 2013.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), October 2010. ISSN 1544-6115. doi: 10.2202/1544-6115.1585.

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. *ArXiv*, abs/2103.00020, 2021.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *ArXiv*, abs/2110.15073, 2021.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Daniel Ting, Ling Huang, and Michael Jordan. An analysis of the convergence of graph laplacians, 2011.

Alexander Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. Diffusion earth mover's distance and distribution embeddings, 2021.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022.

Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Alexander M. Bronstein, I. Oseledets, and Emmanuel Müller. The shape of data: Intrinsic distance for data distributions. *arXiv: Machine Learning*, 2019.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Sathamangalam R Srinivasa Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics*, 20(2): 431–455, 1967.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pp. 555–586, 2008.

Lisa Weijler, Irene Ballester, Guofeng Mei, Tolga Birdal, and Pedro Hermosilla. Latent riemannian flow matching for geometry-grounded 3d foundation models. *ArXiv*, abs/2607.19120, 2026.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin P. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *European Conference on Computer Vision*, 2017.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4540–4549. IEEE, 2019.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17, 2004.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *ArXiv*, abs/2510.11690, 2025a.

Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *ArXiv*, abs/2504.08212, 2025b.

A Riemannian Geometry

A **smooth manifold** M of dimension n is (informally) a space covered by charts to $\mathbb{R}^n$ whose transition maps are smooth. At each point $x \in M$ the **tangent space** T_xM is the n-dimensional vector space of velocities of curves through x, and a choice of local coordinates identifies it with $\mathbb{R}^n$.

A **Riemannian metric** G assigns to each x an inner product G_x on T_xM, varying smoothly in x. In coordinates it is a symmetric positive-definite matrix $G_x \in \mathbb{R}^{n\times n}$, and it measures the length of a tangent vector as $\|\mathbf{v}\|_{G_x}^2 = \mathbf{v}^\top G_x \mathbf{v}$. This is a purely local object: it gives lengths of infinitesimal displacements and says nothing directly about two points a finite distance apart. It generalizes the squared Mahalanobis norm $\mathbf{v}^\top \Sigma^{-1}\mathbf{v}$, except that the matrix is allowed to change from point to point.

Given a smooth map $\psi : \mathcal{Y} \to \mathcal{Z}$ and a metric G on the target, the **pullback metric** on $\mathcal{Y}$ is defined by measuring a tangent vector through its image:

$$(\psi^* G)_y(\mathbf{v}, \mathbf{v}) \;=\; G_{\psi(y)}(d\psi_y \mathbf{v},\, d\psi_y \mathbf{v}) \;=\; \mathbf{v}^\top J_\psi(y)^\top G_{\psi(y)}\, J_\psi(y)\, \mathbf{v}, \tag{9}$$

where $J_\psi(y)$ is the Jacobian. This is positive semidefinite rather than positive definite whenever $d\psi_y$ has a kernel, and those null directions are the displacements ψ cannot see: $\mathbf{v}$ has zero pullback length exactly when $J_\psi(y)\mathbf{v} = 0$. With G the identity the pullback is $J^\top J$, the Gauss-Newton matrix of ψ. We use this construction in Appendix C.

The **length** of a curve $\gamma : [0,1] \to M$ is $\int_0^1 \|\dot\gamma(s)\|_{G_{\gamma(s)}} ds$, and the **geodesic distance** $d_{\mathrm{geo}}(x, y)$ is the infimum of lengths over curves joining x and y. This is how a local metric is converted into a global distance, and it requires solving a variational problem for each pair.

Finally, the **Laplace-Beltrami operator** $\Delta = -\operatorname{div}\operatorname{grad}$ is the generalization of the Euclidean Laplacian to M, defined using the metric to form the gradient and the volume form to form the divergence. It is self-adjoint and non-negative on $L^2(M)$, and on a compact manifold its spectrum is discrete (Appendix B). On $\mathbb{R}^n$ it is $-\sum_i \partial^2/\partial x_i^2$, and in general $\Delta f(x)$ is, up to a positive constant, the limit of $\big(f(x) - \text{mean of } f \text{ over a small geodesic ball around } x\big)$ divided by the squared radius. It is also, up to a factor of two, the negative generator of Brownian motion on M, which is the link to the heat kernel.

B Heat Kernels on Manifolds

Existence and the spectral expansion. Let M be a compact, connected Riemannian manifold without boundary and $\Delta = -\operatorname{div}\operatorname{grad}$ its Laplace-Beltrami operator. Δ is self-adjoint and non-negative on $L^2(M)$, and on a compact manifold its resolvent is compact, so its spectrum is discrete: $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \cdots \to \infty$, each eigenvalue of finite multiplicity, with eigenfunctions $\{\varphi_k\}$ forming an orthonormal basis of $L^2(M)$. Applying $e^{-t\,\cdot}$ to this spectrum gives the heat kernel

$$p_t(x, y) = \sum_{k\ge 0} e^{-t\lambda_k} \varphi_k(x)\varphi_k(y), \tag{10}$$

the series converging uniformly on $M \times M$ for every $t > 0$ since λ_k grows like $k^{2/n}$ (Weyl's law) while $\|\varphi_k\|_\infty$ grows at most polynomially in λ_k. In general the eigenfunctions are the analog of Fourier modes adapted to the geometry of M, ordered from smooth to oscillatory.

Basic properties. The heat kernel is symmetric, $p_t(x, y) = p_t(y, x)$, and strictly positive for all $t > 0$ by the strong maximum principle: heat released anywhere reaches everywhere immediately. It conserves mass, $\int_M p_t(x, y)\, dy = 1$, so $p_t(x, \cdot)$ is a probability density, and more specifically the transition density of Brownian motion on M started at x and observed at time $2t$. It satisfies the semigroup property

$$p_{s+t}(x, y) = \int_M p_s(x, z)\, p_t(z, y)\, dz, \tag{11}$$

which follows from Equation 10 and the orthonormality of the φ_k, and $p_t(x, \cdot) \to \delta_x$ weakly as $t \to 0$. In $\mathbb{R}^n$ it is the Gaussian density $(4\pi t)^{-n/2} \exp(-\|x - y\|^2/4t)$. On a manifold it plays

the same role, and the semigroup property is the Chapman-Kolmogorov equation of that Markov process.

The feature map. Equation 10 is a sum of non-negative coefficients times outer products, so p_t is a positive semidefinite kernel for every $t > 0$. Setting $\Phi_t(x) = (e^{-t\lambda_k/2}\varphi_k(x))_{k\geq 0}$ gives $p_t(x, y) = \langle \Phi_t(x), \Phi_t(y)\rangle$, so Φ_t is the spectral square root of the heat semigroup: explicitly, it is the Mercer decomposition of the kernel p_t. The Euclidean distance between these embeddings is the distance induced by the kernel p_t,

$$\|\Phi_t(x) - \Phi_t(y)\|_2^2 \;=\; p_t(x,x) + p_t(y,y) - 2\,p_t(x,y) \;=\; \sum_{k\geq 1} e^{-t\lambda_k}\left(\varphi_k(x) - \varphi_k(y)\right)^2, \quad (12)$$

where the $k = 0$ term drops out because φ_0 is constant. The diffusion distance of Coifman & Lafon (2006a) at time s is $D_s(x, y) = \|p_s(x,\cdot) - p_s(y,\cdot)\|_{L^2(M)}$, and by the semigroup property and orthonormality of the φ_k it equals $\left(\sum_{k\geq 1} e^{-2s\lambda_k}(\varphi_k(x)-\varphi_k(y))^2\right)^{1/2}$. The continuous embedding Φ_t therefore realizes the diffusion distance at half the time (in the discrete construction the factor μ_k adds one step of the walk, Section 4),

$$\|\Phi_t(x) - \Phi_t(y)\|_2 \;=\; D_{t/2}(x, y), \quad (13)$$

which is why a statistic computed on Φ_t is a statistic on an ordinary Euclidean space, whatever the geometry underneath. The diffusion distance compares the transition densities of the two starting points in L^2, so it is small when the two points reach the same regions with the same probabilities, which requires many short paths between them rather than one.

What t controls. The factor $e^{-t\lambda_k}$ suppresses high-frequency modes, so t acts as a resolution parameter: $e^{-t\Delta}$ is a low-pass filter on M, and in $\mathbb{R}^n$ it is convolution with a Gaussian of variance $2t$ per coordinate. As $t \to 0$ the kernel concentrates and recovers local geometry. Varadhan's formula (Varadhan, 1967) states that $-4t \log p_t(x, y) \to d_{\text{geo}}(x, y)^2$. At intermediate t progressively coarser structure dominates. As $t \to \infty$ only $\lambda_0 = 0$ survives, $p_t \to 1/\operatorname{Vol}(M)$, and the rate of approach is governed by the spectral gap λ_1. Using several t and concatenating the resulting embeddings preserves positive semidefiniteness, since a sum of positive semidefinite kernels is positive semidefinite.

Why not geodesic distance. There are two reasons not to use the geodesic distance instead. First, geodesic distance is determined by a single shortest path, whereas the heat kernel aggregates over all paths of duration t, so it summarizes the neighborhood rather than one route through it. It is correspondingly stable under the addition or removal of a single edge, which a shortest path is not. Second, Feragen et al. (2015) proved that a Gaussian kernel built on the geodesic distance of a complete Riemannian manifold is positive definite for every bandwidth only if the manifold is isometric to a Euclidean space. A divergence built on the geodesic distance would therefore face the same validity problem.

The discrete version. Everything above is the continuous theory, and it motivates the discrete construction of Section 4. There $L = I - D^{-1/2}WD^{-1/2}$ is similar to $I - T$ with $T = D^{-1}W$ the transition matrix of a random walk on the samples, and e^{-tL} is, up to the same change of basis, the transition kernel of the continuous-time version of that walk: the discrete analog of Brownian motion and its heat kernel. The connection is that the normalized graph Laplacian built from Gaussian affinities on n samples converges, after rescaling by the bandwidth, as $n \to \infty$ and the bandwidth shrinks at an appropriate rate, to the Laplace-Beltrami operator of the underlying manifold when the samples are uniform on it (Belkin & Niyogi, 2008; Hein et al., 2007), with the graph eigenvectors converging to the φ_k (Von Luxburg et al., 2008). The raw distance reflects the sampling density as well as the geometry, and the density can be removed by the normalization $\tilde{W}_{ij} \mapsto \tilde{W}_{ij}/(q_i^\alpha q_j^\alpha)$ with $q_i = \sum_j \tilde{W}_{ij}$, where, with a fixed bandwidth, $\alpha = 1$ recovers the Laplace-Beltrami operator independently of how the samples were drawn (Coifman & Lafon, 2006a). Under standard sampling assumptions such graph Laplacians converge to a differential operator on the data manifold whose form depends on the bandwidth rule and the normalization, and the limit depends only on the local behavior of the kernel (Ting et al., 2011; Berry & Sauer, 2014; Calder & Trillos, 2020). With bandwidths that adapt to the local sampling density, as ours do, it is a density-weighted operator (Berry & Harlim, 2014).

None of this convergence is required for our guarantees. Remark 1 holds for any symmetric non-negative W with positive degrees, and Proposition 1 for any distance d, on any finite sample, manifold or not.

C Distances in Feature Space

Evaluation scores are often computed on the features of a pretrained model rather than on the outputs directly, because pooled features are lower-dimensional than the outputs and more robust to small input changes. Our construction applies to features unchanged. The open question is which distance to use, and we propose one derived from a downstream task head.

C.1 Task-Induced Riemannian Metrics

Bond et al. (2026b) introduced a method for learning low-rank approximations of the task-induced Riemannian pullback metric within the feature space of Vision Transformers (ViTs). Here, we review this construction and generalize it to arbitrary feature spaces and backbone architectures. The details are given in Appendix D.

Let $\mathbf{f}^l = \mathrm{vec}(F^l) \in \mathbb{R}^d$ be the vectorized feature tensor at layer l, and let $\psi_l : \mathbb{R}^d \to \mathbb{R}^{d_o}$ be the probe map, the composition of all remaining layers past layer l together with a downstream task head ϕ. We can then produce a task-specific distance on the predictions by equipping $\mathbb{R}^{d_o}$ with a Riemannian metric G. The local sensitivity of the task output to perturbations anywhere across the feature tensor is given by the Jacobian $J_{\psi_l}(\mathbf{f}^l) \in \mathbb{R}^{d_o \times d}$, which induces a pullback Riemannian metric on the vectorized feature space:

$$g(\mathbf{f}^l) = J_{\psi_l}(\mathbf{f}^l)^\top G(\psi_l(\mathbf{f}^l))\, J_{\psi_l}(\mathbf{f}^l) \in \mathbb{R}^{d \times d} \tag{14}$$

For an arbitrary infinitesimal perturbation $\mathbf{v} \in \mathbb{R}^d$, the quadratic form $\mathbf{v}^\top g(\mathbf{f}^l)\mathbf{v} = \|J_{\psi_l}(\mathbf{f}^l)\mathbf{v}\|_G^2$ quantifies the squared rate of change induced in the task space. We take $G = I$ throughout, in which case the quadratic form is $\|J_{\psi_l}(\mathbf{f}^l)\mathbf{v}\|_2^2$. For any $G \succ 0$, directions in $\ker(J_{\psi_l})$ correspond to task-invariant feature variations, while principal eigenvectors of $g(\mathbf{f}^l)$ identify the directions of greatest local sensitivity *of the task output*.

C.2 A Distance Between Samples

The metric $g(\mathbf{f}^l)$ is an infinitesimal object. It assigns a length to a perturbation at $\mathbf{f}^l$ and does not by itself give a distance between two actual samples. The exact finite-distance object it induces is a geodesic length, which would require solving a boundary value problem for every pair, and on a curved space the squared geodesic distance is not of negative type, so Gaussian kernels built on it are not positive definite (Feragen et al., 2015).

We instead use the direct distance, weighted by the task sensitivity at both endpoints:

$$d_{\mathrm{task}}(\mathbf{f}_1, \mathbf{f}_2)^2 \;=\; \tfrac{1}{2}\,(\mathbf{f}_1 - \mathbf{f}_2)^\top\,[\,g(\mathbf{f}_1) + g(\mathbf{f}_2)\,]\,(\mathbf{f}_1 - \mathbf{f}_2) \tag{15}$$

This is symmetric, and $\boldsymbol{\delta} = \mathbf{f}_1 - \mathbf{f}_2$ lives in the feature space. It is also cheap despite the size of g. Writing the rank-r factorization $g(\mathbf{f}) = \sum_{j=1}^r s_j^2 \mathbf{b}_j \mathbf{b}_j^\top$ from the singular value decomposition of $J_{\psi_l}(\mathbf{f})$, with the $\mathbf{b}_j$ the columns of $B(\mathbf{f})$, each of the two quadratic forms in Equation 15 is $\|\mathbf{s}(\mathbf{f}) \odot (B(\mathbf{f})^\top \boldsymbol{\delta})\|_2^2$, which costs $O(dr)$ per pair and never forms the full $d \times d$ matrix. Alternatively, the network introduced by Bond et al. (2026b) could be used to provide a learned parametrization of the metric in feature space.

Equation 15 is a linearization, so it is meaningful only while $\boldsymbol{\delta}$ is small, and it extrapolates badly otherwise. We therefore never use it globally. Instead, we use Euclidean distance to find candidate neighbors and compute the linearized distance only on those candidates, which is all the construction needs. The comparison between distributions is then DED on a graph whose edges carry local task-pullback distances. It is computed in the feature space and is not a distance between the decoded outputs.

Table 4: **Task-induced distances on real features.** Each entry is the Spearman correlation ρ between a score and the degradation severity, with the largest $|z|$ over the severities in parentheses. The *visible* degradation moves the features in a direction the decoder responds to, so every score should detect it. The *irrelevant* degradation moves them in a direction the decoder cannot see, so the correct answer is to report nothing. *const* means the score is exactly constant across severities. The Euclidean distance on the features reports the irrelevant change at $|z| = 202$.

	visible	irrelevant
Euclidean on features	+1.00 (240)	+1.00 (202)
output AbsRel	+1.00 (178)	**const**
output $1 - \delta$-accuracy	+1.00 (200)	**const**

C.3 Feature Spaces Without A Task Head

The above requires a task head ϕ which processes the feature space representation into some given output modality. Often there is none: the encoder is used for retrieval or as a generic backbone, and no decoder has been trained on it. There is then nothing to pull back through, and also no output space in which one could evaluate instead, so the choice of distance has to be made in feature space directly.

The standard approach in this case is to reduce the feature set to one vector. FID and KID compare a mean-pooled activation vector, which identifies any two token sets with the same mean, and CMMD compares the CLS embedding, one vector chosen by the network.

The alternative is to keep every token and compare the sets directly, with Chamfer, Hausdorff or an optimal assignment between the two token sets. On real ViT features the Hausdorff distance between token sets fails negative type, while exact optimal transport passes (Table 15). With no task head there is no output space to evaluate in instead, so for a distance such as Hausdorff our construction is what makes a valid statistic available.

C.4 Evaluation on Real Features

We apply the construction to real ViT features, using layer-11 token features of DINOv2 ViT-B/14 (Oquab et al., 2023) on ImageNet validation images, flattened and reduced to 64 dimensions by PCA. For an exact test we use the simplest task-induced distance, a distance on the decoded outputs, $d(\mathbf{f}_1, \mathbf{f}_2) = d_Z(\psi(\mathbf{f}_1), \psi(\mathbf{f}_2))$, with the decoder $\psi(\mathbf{f}) = \mathrm{softplus}(A\mathbf{f})$ for a random linear map A to 12 outputs, so that $\ker d\psi = \ker A$ exactly. There are two degradations, each at five severities, with $|R| = 300$, $|E| = 150$, generated pools of 150 and $\kappa = 10$. The first, *visible*, moves the features along a direction ψ responds to, so every reasonable distance should detect it. The second, *irrelevant*, moves them along $\ker A$, where the decoded output does not change, so the correct response is none. We compare against the Euclidean distance on the same features, whose pooled version existing scores use.

For d_Z we use the symmetrized AbsRel and $1 - \delta$-accuracy, the fraction of outputs whose ratio exceeds 1.25. The latter is thresholded and violates the triangle inequality, but it is non-negative, symmetric and zero on the diagonal, which is all the construction requires. Both fail negative type on these samples, so energy distance on either carries no guarantee.

Table 4 gives the result. Both output distances track the visible degradation, as the Euclidean distance does, and are exactly constant under the irrelevant one. This holds by construction, since ψ, and hence d, is unchanged along $\ker A$. The Euclidean distance on the same features reports the irrelevant change at $|z| = 202$.

D Task-Induced Riemannian Metrics: Details

Vectorization and architecture independence. Consider layer l of a pretrained network producing a feature tensor $F^l \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$, where the shape $(d_1, \ldots, d_k)$ depends on the model: (N, D) tokens for a ViT, (C, H, W) channels and spatial positions for a CNN, $(|V|, D)$ nodes for a graph network. To analyze metric properties independently of layout conventions, we map F^l to its vec-

torized representation $\mathbf{f}^l = \mathrm{vec}(F^l) \in \mathbb{R}^d$, where $d = \prod_{i=1}^k d_i$ is the total number of scalar feature units at layer l. Every statement below is then a statement about a map $\mathbb{R}^d \to \mathbb{R}^{d_o}$.

Scope of the task metric. Equipping $\mathbb{R}^{d_o}$ with a Riemannian metric G requires the task distance to be induced by a smoothly varying inner product on the output. Distances that are not of this form (a discrete classification loss, a thresholded accuracy ratio) require a separate construction, such as the distance on decoded outputs of Appendix C. A divergence between predicted distributions is of this form to second order, with G the Fisher information. When the output space is Euclidean, $G = I$ and the pullback reduces to $J_{\psi_l}^\top J_{\psi_l}$, which is the case we use throughout.

Rank. Since $J_{\psi_l}(\mathbf{f}^l) \in \mathbb{R}^{d_o \times d}$, the pullback satisfies $\mathrm{rank}(g) \le \min(d_o, d)$, and in practice $d_o \ll d$: a dense depth head at 32×32 output gives $d_o \approx 10^3$ against $d \approx 2 \times 10^5$ for a ViT feature tensor. The metric is therefore positive semidefinite and heavily rank-deficient, with $\ker(J_{\psi_l})$ of codimension at most d_o. Sensitivity typically concentrates far below even that bound, which is what makes a rank-r approximation with $r \ll d_o$ useful. However, further reduction is possible if combined with the latent space of a VAE model Bond et al. (2026a).

Why the metric is computed per sample. g depends on $\mathbf{f}^l$. A single averaged metric $\bar{g} = \mathbb{E}[g(\mathbf{f}^l)]$ would be cheaper and would yield a global Euclidean embedding, but its least-sensitive directions need not be near-null at any individual point, so a perturbation built from $\bar{g}$ can be plainly visible to ψ_l at a particular sample. We therefore compute the factorization at every sample that enters the graph, reference, evaluation and generated alike. This costs r Jacobian-vector and r vector-Jacobian products per sample, after which the quadratic form of Equation 15 costs $O(dr)$ per pair.

E Synthetic Experiments

We use two-dimensional experiments, where the samples and the graph can be drawn, to illustrate the properties of DED and the failures of existing scores. Figures 2 and 3 use the Euclidean distance, and Figures 4 and 5 use modified distances that agree with it on retained edges. All synthetic experiments use $\kappa = 10$, $|R| = 600$, $|E| = 300$ and generated pools of 300 samples, with the other settings of Section 6. The real distribution is a mixture of eight Gaussians with standard deviation 0.45 and means equally spaced on a circle of radius 3 in $\mathbb{R}^2$.

E.1 Controlled Degradations of a Circle of Gaussians

At severity $s \in \{0, 0.25, 0.5, 0.75, 1\}$ we apply four degradations. A *shift* moves the pool by $1.5s$ along the first axis. *Mode drop* keeps only samples with first coordinate above $-3 + 5.7s$ (at $s = 1$, the rightmost mode) and resamples the pool from them. *Mode collapse* keeps n^{1-s} distinct samples and duplicates them (at $s = 1$, a single point). *Off-manifold* replaces a fraction s of the pool with draws from $\mathcal{N}(0, 0.6^2 I)$ at the center of the circle, between the modes. The dispersion of a pool is its mean pairwise distance in the diffusion embedding divided by that of E.

E.2 Comparison with the Fréchet Distance

We draw a Gaussian sample and transform it affinely so that its sample mean and covariance equal those of a fresh sample from the circle of Gaussians, which makes the Fréchet distance between the two samples exactly 0. DED on the Euclidean distance places the Gaussian sample $z = 19.2$ standard deviations above its null, while the fresh circle sample gives $z = 0.0$. This illustrates that energy distance, which is equivalent to an MMD with a characteristic kernel, is not limited to the first two moments.

E.3 Local Invariances

Next, we compare two distances on the circle of Gaussians: d is the Euclidean distance, and

$$d'(x_i, x_j) = \begin{cases} \|x_i - x_j\| & (i,j) \in N \\ F_{ij}\|x_i - x_j\| & \text{otherwise} \end{cases} \tag{16}$$

with $F_{ij} = F_{ji} \sim \mathrm{Uniform}(1, 101)$. This inflates every distance outside the retained edges and leaves the distances on retained edges exact, so d' agrees with d on N and, since inflation cannot

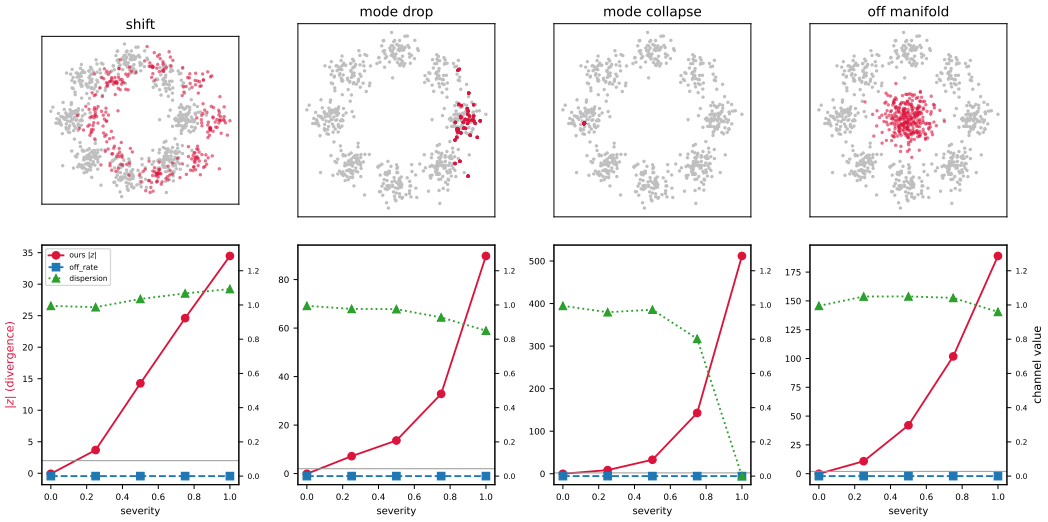

Figure 2: **Controlled degradations of a circle of Gaussians.** The top row shows real samples (gray) and generated samples (red) at maximum severity for each of the four degradations. The bottom row shows DED, reported as $|z|$, together with the off-support rate and the dispersion (blue and green, right axis), as severity increases from 0 to 1. DED rises monotonically on all four degradations. The off-support rate rises when samples leave the region the reference covers, to 0.95 when they are moved off the manifold and to 0.31 under the shift, and stays near zero under mode drop and mode collapse, where the remaining samples are still on the support. Dispersion falls to zero under mode collapse, falls moderately under mode drop and off-manifold movement, and is unchanged by the shift.

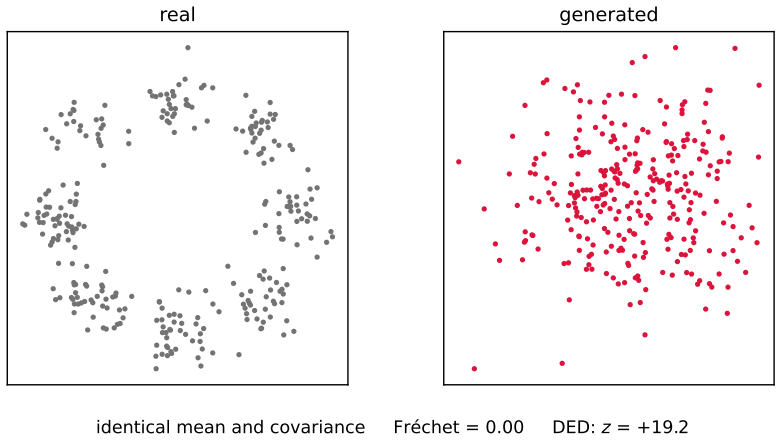

Figure 3: **Matched moments.** The two samples have identical means and covariances but different distributions. The Fréchet distance compares Gaussian fits, so it is 0 and cannot separate them. DED places the Gaussian sample $z = 19.2$ standard deviations above its null, while a fresh sample from the real distribution gives $z = 0.0$.

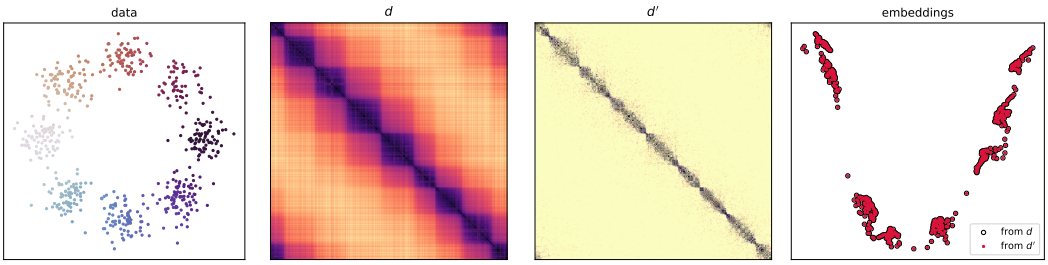

Figure 4: **Locality.** The figure compares the Euclidean distance d with a distance d' that agrees with it on retained edges and is inflated at random elsewhere. The first panel shows the data, colored by angle. The second and third panels show $\log(1+d)$ and $\log(1+d')$ with points sorted by angle: d' agrees with d on a band around the diagonal (the retained edges) and is inflated elsewhere. The fourth panel shows the first two coordinates of the diffusion embedding at t_1, computed from d and from d', and the two embeddings coincide exactly, as Proposition 1 requires.

change a neighbor list, has the same neighbor lists. The corruption destroys negative type: the smallest eigenvalue of the centered distance matrix on R (Appendix M) is 0 up to rounding for d and -3.0×10^3 for d'. Figure 4 shows both distances and the resulting diffusion embeddings.

E.4 Local and Global Failures

Finally, we show that a distance can be exact on every local comparison and still drive energy distance negative, so that it gives no ordering.

We use the same circle of Gaussians. Let $A = \{x \in \mathbb{R}^2 : x_1 > 1.2\}$ (shaded in Figure 5), and let N_{euc} contain every pair in which one point is among the other's κ Euclidean nearest neighbors within the set it is compared with. The distance is

$$d(x,y) = \begin{cases} 61\|x-y\| & (x,y) \notin N_{\text{euc}} \text{ and } x \in A,\ y \in A \\ \|x-y\| & \text{otherwise.} \end{cases} \tag{17}$$

This inflates only pairs that are not Euclidean neighbors, so the neighbor lists of d are the Euclidean ones, its retained edges are exactly N_{euc}, and d is Euclidean on every retained edge. The generated pool is the mode drop of Figure 2, which removes modes from left to right until only the rightmost mode is left at severity 1. In

$$D_d^2 = 2\mathbb{E}d(G,E) - \mathbb{E}d(E,E') - \mathbb{E}d(G,G') \tag{18}$$

a growing fraction of the within-pool pairs are inflated non-neighbor pairs as the pool concentrates in A, so $\mathbb{E}\,d(G,G')$ grows faster than $2\,\mathbb{E}\,d(G,E)$ and the energy distance becomes negative. At severity 1 the pool is a single mode, its within-pool distances are short even after inflation, and the sign flips back.

Replacing energy distance by an MMD with the Gaussian kernel $\exp(-d^2/2\sigma^2)$ on the same distance does not help. When d fails negative type, so does d^2, and by Schoenberg's theorem the kernel is then not positive definite for some bandwidths. At the median-heuristic bandwidth its MMD also turns negative (-0.21 at severity 0.75) and ranks the severities in reverse ($\rho = -0.80$). It fails in the same way at every bandwidth from the median heuristic up to ten times it, and tracks the severity only at a quarter of it or less, where the kernel is effectively zero on all but nearby pairs. DED obtains that locality from the neighbor graph, with no bandwidth to tune: it uses d only on retained edges, where it is exactly Euclidean, and rises monotonically with severity.

The example is constructed to show that the failure can occur. Whether it occurs on real distances is a separate question, answered by Table 15 and Table 2.

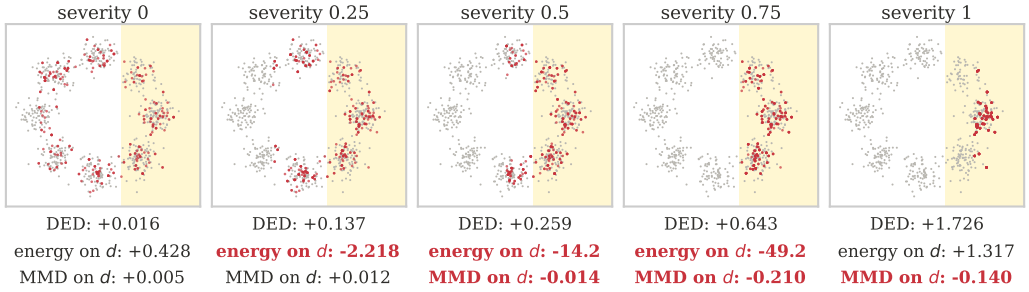

Figure 5: **A locally exact distance that still breaks energy distance.** Each panel shows the reference samples (gray) and the generated pool (red) at one severity, as modes are dropped from left to right, with the inflated region A shaded and three scores printed below. The distance d agrees with Euclidean distance on every κ-nearest-neighbor pair and is inflated by a factor of 61 only on pairs that lie in A and are not neighbors. Energy distance computed directly on d turns *negative* as the pool concentrates in A and is positive again at maximum severity, so it gives no ordering. An MMD with a Gaussian kernel on d, at the median-heuristic bandwidth, also turns negative. DED rises at every step, since it uses d only on the nearest-neighbor pairs, where d is exactly correct.

F Proofs

Proof of Proposition 2. Since $F > 1$, inflating $d(a,b)$ can only move b further down a's list of candidates and vice versa. Since $b \notin \mathcal{N}_\kappa(a)$ and $a \notin \mathcal{N}_\kappa(b)$ already, every neighbor list is unchanged, and d' agrees with d on N because $(a,b) \notin N$. For $P = \frac{1}{2}(\delta_a + \delta_b)$ and $Q = \delta_c$, $\mathbb{E}\, d'(X,Y) = \frac{1}{2}(d(a,c) + d(b,c))$, $\mathbb{E}\, d'(X,X') = \frac{1}{2}Fd(a,b)$ and $\mathbb{E}\, d'(Y,Y') = 0$, so $D^2_{d'}(P,Q) = d(a,c) + d(b,c) - \frac{1}{2}Fd(a,b)$, which is negative for the stated F. □

Proof of the validity remark. (1) Assume every $\deg_i > 0$. For any $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^\top L\mathbf{a} = \sum_{i<j} W_{ij}(a_i/\sqrt{\deg_i} - a_j/\sqrt{\deg_j})^2 \geq 0$, since every $W_{ij} \geq 0$. Hence $L \succeq 0$, and the eigenvalues $e^{-t\lambda_k}$ of $K_t = e^{-tL}$ lie in $(0,1]$. *(2)* With $\Pi = \mathrm{diag}(\pi)$, the normalization of Section 4 gives $\psi_k = \Pi^{-1/2}u_k$, so $\Phi_t\Phi_t^\top = \sum_{k=1}^m \mu_k^2 e^{-t\lambda_k}\psi_k\psi_k^\top = \Pi^{-1/2}\big(\sum_{k=1}^m \mu_k^2 e^{-t\lambda_k}u_k u_k^\top\big)\Pi^{-1/2}$, which is positive semidefinite as a sum of non-negative multiples of outer products. *(3)* The final assertion is the strong negative type of separable Hilbert spaces (Lyons, 2013, Theorem 3.16), applied to $\mathbb{R}^{3m}$. The diffusion embeddings are bounded (Proposition 4), so every pushforward has a finite first moment.

Only symmetry and non-negativity of W are used, with positive degrees. Nothing about d enters beyond what is needed to produce such a W, which is why no condition on d appears in the statement. □

Proof of Proposition 1. The bandwidths satisfy $\sigma_i = \max_{j \in \mathcal{N}_\kappa(i)} d(x_i, x_j)$ and are therefore determined by the stated data. $\tilde{W}$ is supported on N, and its entries there are functions of $d|_N$ and σ alone. Finally W, L and K_t are functions of $\tilde{W}$. The eigenvectors, and hence Φ_t, are determined by L up to an orthogonal change of basis within each eigenspace, which multiplies every diffusion embedding by the same orthogonal matrix and leaves every distance between them, and hence DED, unchanged, provided $\lambda_m < \lambda_{m+1}$.

The same argument gives the extension. Write $N_z = \{z\} \times \mathcal{N}_\kappa(z)$. The bandwidth $\sigma_z = \max_{j \in \mathcal{N}_\kappa(z)} d(z, x_j)$ is a function of $d|_{N_z}$, and $\tilde{W}_{zj}$ is supported on $\mathcal{N}_\kappa(z)$, where its entries are functions of $d|_{N_z}$ and the reference bandwidths σ_j. The remaining quantities follow in turn: $q_z = \sum_j \tilde{W}_{zj}$, then $W_{zj} = \tilde{W}_{zj}(q_z q_j)^{-\alpha}$, then $\deg_z$, and then every $p(z,j) = W_{zj}/\deg_z$ in Equation 6. The reference quantities q_j, λ_k and ψ_k are held fixed throughout, so $\hat{\Phi}_t(z) = \sum_j p(z,j)\,(e^{-t\lambda_k/2}\psi_k(j))_{k=1}^m$ depends on d through $d|_N$ and $d|_{N_z}$ alone. The same holds for the

off-support statistic: $h(z)$ uses d only on N_z and the bandwidths σ_j, and each reference value of h, hence τ, uses d only on N. Nothing about d elsewhere enters, which is why the restriction in Equation 6 matters for our claim and not just for cost. □

What holds without injectivity. Energy distance on $\mathbb{R}^{3m}$ is, up to a factor of two, the squared MMD of the distance kernel, which is characteristic on $\mathbb{R}^{3m}$, so DED compares the entire distributions of the diffusion embeddings, not finitely many moments, whether or not $\hat{\Phi}$ is injective. On the reference set, with all $n-1$ nontrivial modes retained and every $\mu_k \neq 0$, Φ_t separates the reference points: the matrix $[\mathbf{1}, \psi_1, \ldots, \psi_{n-1}]$ is invertible, so no two of its rows coincide, and multiplying column k by the nonzero factor $\mu_k e^{-t\lambda_k/2}$ preserves this. Truncating to m modes, which we do for efficiency (the one-step placement needs no spectral cut), can make distinct points collide. Off the reference, injectivity also fails through the placement rule (Section 5.1).

The three conditions for a divergence. A divergence is usually required to be non-negative, to vanish exactly when its two arguments coincide, and to be locally positive definite, meaning that for nearby arguments its leading term is a positive-definite quadratic form in their difference. Energy distance satisfies all three exactly. Expanding the right-hand side shows that

$$D^2(P,Q) = -\iint \|x-y\|_2 \, d(P-Q)(x) \, d(P-Q)(y), \tag{19}$$

so D^2 is a quadratic form in $P-Q$ with no higher-order terms. By strong negative type this form is strictly positive for every nonzero signed measure of total mass zero with finite first moment (Lyons, 2013), which gives the three conditions at once. For DED the form is applied to $\hat{\Phi}_{\#}P - \hat{\Phi}_{\#}Q$, and the diffusion embeddings are bounded (Proposition 4), so DED satisfies the three conditions as a function of the distributions of the diffusion embeddings. As a function of $P-Q$ itself the form is positive semidefinite, and it is degenerate exactly along the differences that $\hat{\Phi}$ maps to zero, which is the injectivity caveat above.

Proof of Proposition 3. Size. $\hat{\Phi}$ and τ are functions of R alone, and E and G are drawn independently of R. Conditional on R, each sample z of $E \cup G$ is mapped to the pair $w(z) = (\hat{\Phi}(z), \mathbf{1}[h(z) < \tau])$, and $h(z)$ is computed against R only. Under $P = Q$ the $n_E + n_G$ pairs are therefore i.i.d., hence exchangeable. Every statistic used here, the U-statistic, the difference of off-support rates, their permutation p-values and the minimum of those, is a function of the pairs and of how they are split. Let S_0 be the value of $\min(p_D, p_O)$ at the observed split and $S_1, \ldots, S_B$ its values at B uniformly random re-splits, with p_D and p_O at each split computed against all $B+1$ splits. By exchangeability $S_0, \ldots, S_B$ are exchangeable, so the rank of S_0 among them is uniform up to ties, and $p = \#\{b : S_b \leq S_0\}/(B+1)$ satisfies $\Pr(p \leq a) \leq a$ for every a (Phipson & Smyth, 2010).

Consistency. For every $s \in [0,1]$, the number of splits b with $p_D(b) \leq s$ is at most $s(B+1)$: if b^* has the smallest statistic among them, every one of them has a statistic at least as large as that of b^*, and there are $p_D(b^*)(B+1) \leq s(B+1)$ such splits. The same holds for p_O. Hence $p \leq 2S_0 = 2\min(p_D, p_O)$ at the observed split. Fix R, so that $\hat{\Phi}$ and τ are fixed. The diffusion embeddings lie in the convex hull of the reference points' vectors v_j (Proposition 4), so they are bounded, and the energy statistic converges to $D^2(\hat{\Phi}_{\#}P, \hat{\Phi}_{\#}Q) > 0$ when the pushforwards differ (Székely & Rizzo, 2013), while its permutation distribution concentrates at 0. The same holds for the one-sided difference of off-support rates when $\Pr_Q[h(Z) < \tau] > \Pr_P[h(X) < \tau]$. In either case the observed split has the strictly largest value of that statistic with probability tending to one as $n_E, n_G \to \infty$, so that marginal p-value tends to $1/(B+1)$ and $\Pr(p \leq a) \to 1$ for every level $a \geq 2/(B+1)$ (≈ 0.01 at $B = 200$). □

G ALGORITHM

Algorithm 1 Diffusion Energy Divergence

Require: reference set R, evaluation set E, generated sets $G_1, \dots, G_C$, distance d, neighbors κ, density exponent α, number of modes m, number of permutations B

1: **Neighborhoods.** For each $x_i \in R$ compute $\mathcal{N}_\kappa(i)$ and $\sigma_i = \max_{j \in \mathcal{N}_\kappa(i)} d(x_i, x_j)$, and let N be the symmetrized retained edge set.

2: **Local affinities.** $\tilde{W}_{ij} = \exp(-d(x_i, x_j)^2/(\sigma_i \sigma_j))$ for $(i, j) \in N$, else 0.

3: **Density normalization.** $q_i = \sum_j \tilde{W}_{ij}$, floored at 0.05 times its median, and $W_{ij} = \tilde{W}_{ij}(q_i q_j)^{-\alpha}$.

4: **Laplacian and spectrum.** $L = I - D^{-1/2} W D^{-1/2}$, then eigendecompose $L = \sum_k \lambda_k u_k u_k^\top$.

5: **Retain modes.** Keep the m smallest nonzero eigenvalues. Set $\psi_k = u_k/u_0$, where $u_0 = (\deg/\mathrm{vol})^{1/2}$ and $\mathrm{vol} = \sum_i \deg_i$, and $\mu_k = 1 - \lambda_k$. Set $t_1 = 1/\lambda_{\max}$, $t_3 = 1/\lambda_{\min}$ and $t_2 = \sqrt{t_1 t_3}$ over the retained eigenvalues.

6: **Diffusion embedding.** $\Phi_t(x_i) = (\mu_k e^{-t\lambda_k/2} \psi_k(i))_k$, concatenated over the three times.

7: **Threshold.** τ = the 1% quantile of $h(x_i)$ over R, each computed against the other $n - 1$ reference points.

8: **Extension.** For each $z \in E \cup G_1 \cup \cdots \cup G_C$, form $\mathcal{N}_\kappa(z) \subseteq R$, compute $\tilde{W}_{zj}$ on it only, and apply Equation 6 with q_j, λ_k, ψ_k held fixed at their reference values. Flag z off-support if $h(z) < \tau$.

9: **Score.** For each c, report the U-statistic of Equation 8 between $\hat{\Phi}(E)$ and $\hat{\Phi}(G_c)$, the off-support rate, and the combined p-value of Proposition 3 over B permutations.

G.1 DETAILS OF THE PROCEDURE

Connectivity. The multiplicity of the eigenvalue 0 of L is the number of connected components of the reference graph. If the graph is disconnected, the coarse scales of the diffusion embedding encode component membership rather than geometry, and κ should be increased. For some data and distances no reasonable κ connects the graph. Since every generator is compared on the same graph, scores remain comparable, but the coarse scales should then be read with care.

Diffusion times. The finest scale the retained modes resolve is $t_1 = 1/\lambda_{\max}$ and the coarsest is $t_3 = 1/\lambda_{\min}$. The three times of Section 5 are spaced geometrically between them, and the three blocks of the embedding are concatenated without rescaling. A sum of positive semidefinite kernels is positive semidefinite, so validity is unaffected. The divergence can also be reported as a function of t, which shows at which scale two distributions differ, together with a single test that takes the maximum standardized divergence over a grid of times and calibrates it by permutation, as MMDAgg does over kernel bandwidths (Schrab et al., 2021). Appendix I varies κ, the number of modes, the number of diffusion times, α and $|R|$, and redraws the reference set.

Placement. Adding evaluation or generated samples to the graph and recomputing the spectrum would move the diffusion embedding of every reference point, and two generators scored against the same data would no longer share one embedding, which is why samples are placed rather than added. Placing z with its distances to every reference point, rather than to its κ nearest, would make the score depend on the values of d between z and distant reference points, which would break Proposition 1.

Because $p(z, \cdot)$ is a probability vector, $\hat{\Phi}_t(z)$ lies in the convex hull of its neighbors' vectors v_j, and the placement is Lipschitz in the neighbor weights with no amplification (Proposition 4). The standard Nyström rule (Bengio et al., 2003; Coifman & Lafon, 2006b) instead divides by μ_k, which is why it needs a spectral cut. The density factor q_i is floored at 0.05 times its median over R, which guards against nearly isolated points. The floor depends on d only through $d|_N$, so Proposition 1 is unaffected.

The estimator. Being unbiased, Equation 8 can be slightly negative when $P = Q$, while the population value cannot. The permutation test pools $\hat{\Phi}(E)$ and $\hat{\Phi}(G)$, re-splits them at sizes (n_E, n_G)

and recomputes Equation 8. It is exact under $P = Q$ because the pooled diffusion embeddings are then exchangeable.

Off-support samples. The log affinity mass $h(z)$ decreases without bound as z moves away from the reference, and it is measured in units of the local bandwidths, so it adapts to regions of different sampling density. Multiplying all of $\tilde{W}_{z\cdot}$ by a common factor leaves $\hat{\Phi}_t(z)$ unchanged, which is why the diffusion embedding alone cannot see it.

Every sample is kept and placed. The threshold τ is a quantile of the reference points' own affinity masses, each computed against the other $n-1$ points, so it adapts to the distance and to the density of the reference without a tuned constant, and real samples drawn like the reference fall below it about 1% of the time. The rate is tested jointly with the divergence (Proposition 3) and does not enter the diffusion embedding. Appending h to the embedding as one more dimension would fold the two statistics into one divergence. When a whole generator lies off the reference support, as 16 of the 20 VBench generators do relative to MSR-VTT footage (Appendix L.7), that coordinate then dominates the divergence and hides changes within the generator's own pool, so we keep the two separate and combine only their p-values.

Other feature spaces. The construction is indifferent to where d comes from, so it applies unchanged in the feature space of a pretrained encoder. There the open question is which distance to use, and Appendix C gives one such option.

G.2 Derivation of the Nyström Extension

The eigenvectors of L can be viewed as samples of eigenfunctions of a diffusion operator, which is what allows them to be evaluated away from the reference points. Let $T = D^{-1}W$ be the random walk of Section 4, with entries $p(i,j) = W_{ij}/\deg_i$, right eigenvectors ψ_k and eigenvalues $\mu_k = 1 - \lambda_k$ (T is similar to $D^{-1/2}WD^{-1/2} = I - L$). Coordinatewise, $T\psi_k = \mu_k\psi_k$ gives

$$\mu_k\,\psi_k(i) = \sum_j p(i,j)\,\psi_k(j). \tag{20}$$

At a reference point this holds by definition. At a new point z, attached through its own κ nearest reference points as in Section 5 (so $p(z,j) = 0$ for $j \notin \mathcal{N}_\kappa(z)$), we can then choose to use it as the definition of $\psi_k(z)$, which gives the Nyström extension $\psi_k(z) = \mu_k^{-1}\sum_j p(z,j)\,\psi_k(j)$ (Williams & Seeger, 2000; Bengio et al., 2004; 2003). The k-th coordinate of the diffusion embedding is $\mu_k e^{-t\lambda_k/2}\psi_k$ (Equation 5), so at z

$$\mu_k\,e^{-t\lambda_k/2}\,\psi_k(z) = \sum_{j\in\mathcal{N}_\kappa(z)} p(z,j)\,e^{-t\lambda_k/2}\,\psi_k(j), \tag{21}$$

which is Equation 6: the factor μ_k cancels the division, and the placement is one step of the walk from z. With the fitted row $p(i,\cdot)$ of a reference point, the right-hand side returns $\Phi_t(x_i)$ exactly.

Proposition 4 (Stable placement). *Let $v_j = (e^{-t\lambda_k/2}\psi_k(j))_{k=1}^m$ and $V = \max_j \|v_j\|_2$. For any two new samples, or one sample under two distances that agree on N, with random-walk rows p and p', $\|\hat{\Phi}_t(z) - \hat{\Phi}_t(z')\|_2 \le V\,\|p - p'\|_1$. Moreover $\hat{\Phi}_t(z)$ lies in the convex hull of $\{v_j : j \in \mathcal{N}_\kappa(z)\}$.*

Proof. $\hat{\Phi}_t(z) - \hat{\Phi}_t(z') = \sum_j (p_j - p'_j)v_j$, and the triangle inequality gives the bound. The convex-hull statement holds because p is a probability vector. □

The standard Nyström rule, which places z at $(e^{-t\lambda_k/2}\psi_k(z))_k$ with $\psi_k(z) = \mu_k^{-1}\sum_j p(z,j)\psi_k(j)$, divides by μ_k, so its bound carries an extra factor $1/\min_k|\mu_k|$, which is unbounded unless modes with μ_k near 0 are cut.

H Settings of the Video Experiments

The real clips of Tables 1 and 3 come from RealCam-Vid (Zheng et al., 2025b). Each clip has 14 frames at 256×256 pixels, processed at 7 frames per second from a window at the center of the

source clip. The motion profile of a clip is the RAFT (Teed & Deng, 2020) optical flow between consecutive frames, averaged over each cell of a 4×4 grid, with every coordinate standardized by its mean and standard deviation over the reference set. The distance between two clips is dynamic time warping between their profiles, restricted to a band of 4 frames around the diagonal.

Table 1 uses $|R| = 200$, $|E| = 120$ and degraded pools of 120 clips, at the severities $s \in \{0, 0.25, 0.5, 0.75, 1\}$. Color jitter multiplies each frame by its own gain drawn from $\mathcal{N}(1, (0.25s)^2)$ and adds its own offset drawn from $\mathcal{N}(0, (0.15s)^2)$. Shuffle permutes a fraction s of the frame positions among themselves. Local reorder permutes the frames inside each window of 4 consecutive frames with probability s. Freeze replaces a fraction s of the frames at the end of the clip by its first frame. Drop-hold replaces a fraction s of the frames by the frame before them. Temporal blur applies a Gaussian filter along time with a standard deviation of $3s$ frames. FVD uses the Kinetics-400 S3D network (Xie et al., 2017) of torchvision in place of the I3D network of the original FVD, or the self-supervised VideoMAE-Base network (Tong et al., 2022), and its z comes from 100 random re-splits of the real clips.

Table 3 uses $|R| = 200$ and $|E| = 60$. Stable Video Diffusion (Blattmann et al., 2023) generates 60 clips per setting, each with 14 frames at 256×256 pixels, from the first frames of RealCam-Vid clips. The sampler steps are 1, 2, 4, 8, 15 and 25 at motion bucket 127, and the motion buckets are 10, 60, 127 and 200 at 25 steps.

Appendix L.6 uses 600 MSR-VTT clips (Xu et al., 2016) processed at 8 frames per second. The sets R (200 clips) and E (100 clips) are drawn from them without overlap, with weights that match the distribution of motion magnitude in the generated clips.

I Sensitivity to Hyperparameters and to the Reference Set

Five choices in the construction could reasonably be made differently: the number of neighbors κ, the number of retained modes m, the number of diffusion times, the density exponent α and the size of the reference set $|R|$. We vary each one alone, holding the others at the defaults of Section 6, on the three NYU depth degradations of Table 2: $\kappa \in \{4, 8, 16, 32\}$, $m \in \{16, 32, 64, 128\}$, $\{1, 2, 3, 5\}$ diffusion times, $\alpha \in \{0, 0.5, 1\}$ and $|R| \in \{100, 200, 300\}$. The sweep uses five severities per degradation at the strength of Table 2, so ρ moves in steps of 0.10. Table 5 reports, for each choice, the largest change in Spearman ρ across its values and the smallest $\max|z|$ reached on any degradation.

Table 5: **Sensitivity to the hyperparameters, on the NYU depth degradations.** Each hyperparameter is varied alone over the values listed in the text, with the others at their defaults. The first three columns give, for each degradation, the largest change in the Spearman correlation ρ between any two of its values, so 0.00 means the ranking never changed. The last column gives the smallest $\max|z|$ reached over its values and the three degradations, so a large value means every setting still detects every degradation clearly.

	largest change in ρ			
varied	collapse	corrupt	scale	min $\max\|z\|$
κ	0.00	0.10	0.00	8
modes m	0.00	0.10	0.00	10
diffusion times	0.00	**0.40**	0.00	10
α	0.00	0.00	0.00	12
$\|R\|$	0.00	0.10	0.00	12

Only the number of diffusion times matters: a single time gives $\rho = +0.60$ on the corruption degradation, while two or more yield $+1.00$ over these five severities, which is why we use three. The reference set can be small. With $|R| = 100$ every degradation still gives $+1.00$, with $\max|z| \geq 27$.

A weaker corruption. At a third of the corruption strength of Table 2, the energy test never rejects ($p \geq 0.90$ at every severity): the corrupted patches are placed where clean ones with the same neighbors would be. They do leave the support, from 3% at the lowest severity to 97% at the

highest, and the combined test rejects from $q = 1/3$ on ($p = 0.01$). This is the case the combined test of Proposition 3 is designed for.

Which reference sample. A fixed reference is what makes scores comparable across generators, so it matters how much the ranking depends on which real samples happen to form R. We rank the 20 VBench generators of Appendix L on the temporal axis, redrawing R and E ten times from each of two real corpora, MSR-VTT and the real clips of VideoFeedback (He et al., 2024), disjoint and without replacement, and compare the rankings by Kendall's τ (Table 6).

Table 6: **How much the ranking of the 20 VBench generators depends on the reference set.** Each entry is Kendall's τ between two rankings, as mean $\pm$ s.d. over pairs of draws, where $\tau = 1$ means identical rankings. The first two rows redraw R and E from the same real corpus, and the last row compares rankings made against the two different corpora.

	τ
redraw R, E within MSR-VTT	0.77 ± 0.07
redraw R, E within VideoFeedback	0.80 ± 0.07
MSR-VTT against VideoFeedback	0.46 ± 0.08

Redrawing the reference from one corpus moves the ranking modestly, and changing the corpus moves it more. The ends of the ranking agree across corpora (Latte-1, Lavie and Mochi near the top, AnimateDiff-V1, AnimateLCM and EasyAnimate near the bottom), and so does the disagreement with VBench: Kendall's τ between DED and VBench's subject consistency is -0.45 on MSR-VTT and -0.54 on VideoFeedback. The models that move most, the two CogVideoX variants, rank 8th and 10th against MSR-VTT and 17th and 18th against VideoFeedback. A ranking is therefore relative to its reference, and so must be specified when using this approach, as discussed in Appendix L.8. Additionally, if instead the symmetric-normalization extension divides by μ_k, the cross-corpus agreement falls to $\tau = 0.15 \pm 0.10$, consistent with the amplification that the one-step placement avoids (Proposition 4).

J Integral Probability Metrics and MMD

Intuitively, the energy distance between P and Q is 0 when the average distance between a sample from P and a sample from Q equals the mean of the average distance between two samples from P and the average distance between two samples from Q.

A Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ is a Hilbert space of functions in which evaluation at a point is a continuous linear functional, so that for every x there is a $k_x \in \mathcal{H}_k$ with $f(x) = \langle f, k_x \rangle$ for all f. The kernel $k(x, y) = \langle k_x, k_y \rangle$ determines the space entirely. A distribution P is then represented by its mean embedding $\mu_P = \mathbb{E}_{X\sim P}[k_X] \in \mathcal{H}_k$, and MMD is the distance $\|\mu_P - \mu_Q\|_{\mathcal{H}_k}$ between two such embeddings. A kernel is *characteristic* exactly when $P \mapsto \mu_P$ is injective, which is what makes this distance zero only when $P = Q$.

MMD belongs to the family of Integral Probability Metrics (IPMs). Given a class of bounded continuous functions $\mathcal{F}$, the IPM distance between two distributions P, Q is defined as:

$$d_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]| \tag{22}$$

Here, the choice of class $\mathcal{F}$ determines the choice of statistical distance. The relevant class for our purposes is the class $\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$ for some RKHS $\mathcal{H}_k$, which represents the unit ball of functions in that RKHS.

Given the kernel k for the RKHS, we can explicitly write the formula for MMD as:

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}_{X,X'\sim P}[k(X, X')] - 2\mathbb{E}_{X\sim P,Y\sim Q}[k(X, Y)] + \mathbb{E}_{Y,Y'\sim Q}[k(Y, Y')] \tag{23}$$

The standard unbiased estimator from finite samples is the U-statistic

$$\widehat{\mathrm{MMD}}^2_U = \frac{1}{n_P(n_P-1)} \sum_{i\neq i'} k(x_i, x_{i'}) + \frac{1}{n_Q(n_Q-1)} \sum_{j\neq j'} k(y_j, y_{j'}) - \frac{2}{n_P n_Q} \sum_{i,j} k(x_i, y_j) \tag{24}$$

K AUDIO

Audio has the same split as video, under different names. Fréchet Audio Distance is computed on embeddings of short windows, so it is dominated by *timbre*, the spectral content of the sound, while the temporal structure of the signal, meaning its rhythm, onset placement and articulation, is what a listener actually notices. Timbre is to audio what appearance is to video, and rhythm is to audio what motion is to video, and we find analyze the same kind of failure.

In video, optical flow supplies the appearance invariance for free, since flow does not see color. However, audio has no such easy solution. A static equalizer, or a different microphone, multiplies the spectrum by a fixed gain per frequency band, which adds a fixed offset to every frame of the log-mel spectrogram. We therefore difference the log-mel spectrogram across time, which cancels that offset exactly, and summarize each difference by its L^1 and L^2 norms, its mean and the change in its L^1 norm. The profile thus describes how the spectrum moves rather than what it contains, and the distance between two clips is DTW on that profile. As in video, this distance satisfies neither the triangle inequality nor negative type, so it cannot be used with energy distance directly.

We corrupt LibriSpeech (Panayotov et al., 2015) clips along one useless axis and five temporal ones. The useless axis is an equalizer applied to the whole clip, which changes the timbre and leaves the timing untouched, so the correct response to it is no response at all. The temporal axes are frame shuffling, local reordering, dropping and holding frames, blurring across frames, and changing the speed. Selectivity is the mean ρ over the five temporal corruptions minus the ρ on the nuisance axis, with a constant score counted as 0, and measures whether a change in the score can be attributed to timing. Both baselines use the same order-free timbre features: the mean and standard deviation of each log-mel band over the clip.

Table 7: **Audio, controlled degradations of LibriSpeech clips.** Each entry is the Spearman correlation ρ between a score and the corruption strength. The equalizer (EQ) changes only the timbre and leaves the timing untouched, so the correct response to it is none, while the other five corruptions change the timing and should be detected. *const* means the score is exactly constant across strengths. Selectivity is the mean ρ over the five timing corruptions minus the ρ on the equalizer, counting a constant score as 0, so $+1.00$ is the best possible value.

	nuisance	temporal					
	EQ	shuffle	reorder	drop-hold	blur	speed	selectivity
Baselines							
Fréchet (timbre features)	+1.00	const	const	+0.90	+1.00	+1.00	-0.42
energy (timbre features)	+1.00	const	const	+0.90	+1.00	+1.00	-0.42
Ours							
DED	**const**	**+1.00**	**+1.00**	**+1.00**	**+1.00**	**+1.00**	**+1.00**

DED is exactly constant on the equalizer axis and rises monotonically with all five temporal corruptions (Table 7), with $|z|$ between 39 and 118 at the highest severity. The Fréchet baseline responds to the equalizer as strongly as to any temporal corruption, which gives it a selectivity of -0.42: a change in its value cannot be attributed to timing rather than to timbre. The same happens in video, where FVD responds to color jitter.

Both baselines are also constant under frame shuffling and local reordering, the two corruptions that change timing while leaving the multiset of frames untouched. A statistic computed on features pooled over frames cannot see a reordering of those frames, which is why we construct the temporal profile the way we did.

L WHAT CURRENT VIDEO BENCHMARKS MEASURE

The video experiments in Section 6 use controlled corruptions and one generator's own controls. Neither tests the benchmarks that the field uses to rank models. This appendix reports what we find when the same questions are asked of VBench (Huang et al., 2023), which is a standard benchmark for text-to-video generation.

VBench scores a generator on sixteen dimensions, five of which make up its temporal quality group. *Subject consistency* is the mean DINO (Caron et al., 2021) feature similarity across frames, *background consistency* the same with CLIP (Radford et al., 2021) features, *temporal flickering* the mean absolute difference between frames of static clips, *motion smoothness* a score from the motion priors of a frame-interpolation model, and *dynamic degree* an estimate of how much movement a clip contains, obtained with RAFT (Teed & Deng, 2020) optical flow. The substitute and leaderboard comparisons below use VBench's released code in its custom-input mode. Where we need per-clip control over frame order (Appendix L.6) we recompute subject consistency (with DINOv2 (Oquab et al., 2023) features), flickering and a continuous flow magnitude ourselves. Across the 14 generators where both are available, our subject consistency agrees with VBench's at $\rho = +0.92$.

Important to note, all five dimensions we study for VBench are **intrinsic**: each is a scalar computed inside a single video, then averaged over the generated pool, with real data never being used as a comparison. Such an intrinsic per-sample scalar is not a divergence, and cannot say whether two sets of generated videos differ.

VBench is explicit that this has consequences. Its authors introduce *dynamic degree* precisely because "a completely static video can score well in the aforementioned temporal quality dimensions". The remedy is a counterweight: a second intrinsic scalar whose optimum sits at the opposite extreme. What follows asks how far the problem extends and whether a counterweight addresses it.

L.1 Real video is not optimal under these scores

We take 300 real MSR-VTT clips (Xu et al., 2016) and build four degenerate substitutes. Three are built from *each* clip: its first frame repeated, a linear blend from its first frame to its last, and its first frame plus small noise. The fourth is fresh noise on every frame. The first three comparisons are therefore paired with the clip they came from, so a difference in content cannot produce the result.

Table 8: **Three temporal dimensions, from VBench's released code, on 300 real clips and on four substitutes.** All three are higher-is-better in VBench's convention. Bold marks each dimension's best value, and none of them is real video. Flickering is computed without the static-clip filter. The win rate of a substitute is the fraction of clips on which it scores higher than the real clip it is paired with, given here with 95% bootstrap confidence intervals over clips. On subject consistency the win rates are 0.997 for the frozen frame, 0.967 [0.943, 0.983] for still + noise and 0.897 [0.860, 0.930] for fresh noise. On flickering they are 1.000 for the frozen frame, 0.990 [0.977, 1.000] for the crossfade and 0.907 [0.870, 0.937] for still + noise. On dynamic degree fresh noise wins on 0.470 of clips, ties on the rest and never loses.

	subj. consistency	flickering	dynamic degree	beats real on
real video	0.908	0.975	0.530	
frozen frame	**1.000**	**1.000**	0.000	consistency, flickering
still + noise	0.998	0.997	0.000	consistency, flickering
crossfade	0.918	0.989	0.003	flickering
fresh noise / frame	0.993	0.762	**1.000**	consistency, dynamic

A linear blend between two of a clip's own frames contains no camera or subject motion, yet it is not static, so it lies outside the *completely static* case VBench's authors anticipate. It beats the clip it was built from on flickering in 99.0% of cases, and on subject consistency it is level with it (49.7%).

Every clip is stored as H.264, which adds error of its own: $1.9/255$ per pixel on the real clips. On the still-plus-noise substitute, whose injected noise is $2/255$, the codec removes about half of it, so we repeat that substitute with lossless encoding on 100 clips. It still beats the real clip on consistency in 98% of cases and on flickering in 74% (95% CI $[65\%, 82\%]$).

The counterweight proposed by the VBench authors has the mirror-image problem. *Dynamic degree* is itself an intrinsic per-video scalar, and a binary one: a clip is either dynamic or not. Fresh noise on every frame is scored dynamic on every clip, so it never loses to real footage, and it wins on all 47% of clips that VBench scores as static. That substitute also beats real video on subject consistency, on 89.7% of clips, because DINO embeds structureless frames close together, so noise is viewed as a highly consistent subject. It fails on flickering.

No single substitute here beats real video on all three dimensions, so the three together do exclude each of these particular non-videos. But every dimension's own optimum lies away from real video, and two of the three dimension *pairs* admit a non-video that is at least as good as real footage on both: consistency with flickering (the frozen and noisy-still substitutes) and consistency with dynamic degree (fresh noise). Only flickering paired with dynamic degree excludes all four, which is unsurprising since those two oppose each other most directly. Adding counterweights narrows the set of non-videos that win, but it does not move any optimum onto the data, and is dependent on the careful selection of weights used to combine these scores. A divergence from real data, in contrast, is minimized by the data itself.

L.2 A REAL GENERATOR IS REWARDED FOR PRODUCING LESS VIDEO

A natural objection to Appendix L.1 is that nobody submits frozen clips to a benchmark. We therefore repeat the question on Stable Video Diffusion (Blattmann et al., 2023) across its own `motion_bucket` conditioning value.

Table 9: **VBench's intrinsic scores and DED across two Stable Video Diffusion settings.** Each entry is the Spearman correlation ρ between a setting and a score, with no sign flip. Subject consistency is higher-is-better. Flickering is reported as the raw mean frame difference and DED is a divergence, so both are lower-is-better. Subject consistency and flickering are recomputed per clip with DINOv2 features. Motion content, the mean step size of the DINOv2 features between frames, checks that the motion setting really changes the amount of motion. DED uses DTW on the DINOv2 drift profile (the temporal axis of Appendix L.7), not the flow distance of Table 3. The number of sampler steps is a genuine quality ordering and serves as a control.

	motion axis	steps axis (control)
subject consistency	-1.00	$+0.77$
temporal flickering	$+1.00$	$+0.94$
motion content (witness)	$+1.00$	$+0.60$
DED	$+1.00$	-0.77

Both intrinsic dimensions improve monotonically as the motion is turned down. The motion bucket does not control quality, only the amount of motion. On these settings SVD's clips stay below real video on subject consistency (0.72 to 0.83, against 0.92) and above it on motion content (0.45 to 0.66, against 0.26), so turning the motion down also moves them toward real video. What the settings show is the direction the intrinsic dimensions reward, whose end point is the frozen clip (Appendix L.1), not that the generator has passed real video.

The *steps* axis is a real quality ordering, and the two intrinsic dimensions disagree there. Subject consistency agrees with it. Temporal flickering instead goes the wrong way at $+0.94$, which means it reports the generator as getting worse as it is given more sampling compute. Motion also rises with steps, so the consistent view is that flickering is largely measuring motion and better sampling produces more of it. DED, meanwhile, measures the quality ordering in the correct direction, at -0.77. On the motion axis it also falls as the motion is turned down, because in this representation the generator's clips move more than real clips at every setting, so less motion brings them closer to the real data. The two kinds of score differ in where they end: DED is minimized at the real distribution, the intrinsic dimensions at a still image. These settings never reach below real video's motion, so they cannot separate the two, while the substitutes of Appendix L.1 do.

L.3 WHAT THE LEADERBOARD IS ORDERED BY

We then ask the same question of published models. We take 20 generators from VBench's released sample set, 100 clips each. Every clip is processed at a fixed 8 frames per second, so that frame spacing is the same for every model and for the real clips. For each model we take VBench's official subject consistency, temporal flickering and dynamic degree, and a continuous measure of motion.

We use two measures of motion. One is VBench's own *dynamic degree*, so that the first correlation below is between two of VBench's temporal dimensions and not between one of theirs and an outside measure of ours. Dynamic degree is binary per clip, so we also report RAFT (Teed & Deng, 2020)

optical flow magnitude normalized by the frame diagonal, which is the same instrument, read continuously. Two details matter. Flow is computed from pixels by a separate network, so it is independent of the DINO features subject consistency uses. A motion statistic built from those same features would be a monotone function of subject consistency itself, since $\|f_{t+1} - f_t\|^2 = 2 - 2\cos(f_t, f_{t+1})$ for normalized features, and correlating the two would show nothing. Normalizing by the diagonal is needed because flow scales with frame size, and these models emit different resolutions, so raw displacement would rank a model higher simply for rendering larger.

Table 10: **Across 20 published generators, VBench's temporal scores are higher for the generators that move less.** The first column gives the Spearman correlation ρ, over models, between each of VBench's official scores and its own dynamic degree, with a 95% bootstrap confidence interval over models. The last column counts the generators that score above real video.

	ρ with dynamic degree	above real video
subject consistency	$\mathbf{-0.46}$ $[-0.76, -0.01]$	15/20
temporal flickering	$\mathbf{-0.57}$ $[-0.84, -0.15]$	5/20
motion smoothness	-0.38 $[-0.79, +0.11]$	17/20

Subject consistency and flickering both rank published generators partly by how little they move, and both intervals exclude zero. Motion smoothness points the same way, but its interval includes zero. Fifteen of the twenty generators score above real video on subject consistency and seventeen on motion smoothness. Our own recomputation, with the continuous flow magnitude in place of the binary dynamic degree, gives $\rho = -0.74$ $[-0.90, -0.41]$ between subject consistency and motion over the same 20 models.

This is a stronger statement than the one VBench's authors make, in two ways. Their caveat concerns a *completely static* video, a case nobody submits. What we measure is that motion is a substantial axis of variation across models that were in fact submitted. Additionally, the failure is not confined to static output: the crossfade of Appendix L.1 changes on every frame, so it lies outside the caveat entirely, and it still outscores the real clip it was built from on flickering.

The counterweight does not resolve this, because *dynamic degree* is itself an intrinsic per-video scalar. Pairing a score whose optimum is a still image with one whose optimum is maximal movement places real video at neither optimum.

L.4 Agreement with human preference

VBench releases pairwise human preferences between videos that different models generated for the same prompt. On the 808 comparisons with a stated preference, among three models, VBench's subject consistency, computed at a fixed 8 frames per second, agrees with the annotators on 0.824 of them. However, a rule that never looks at the videos agrees on 0.860: given a pair, it predicts whichever model wins more often overall, refitted on held-out prompts so that it never sees the pairs it judges. The two make the same prediction on 77% of pairs. On the 189 pairs where they differ, subject consistency is right 0.423 of the time (95% CI $[0.339, 0.503]$, resampling prompts), no better than chance.

Table 11: **Agreement with human preference on VBench's pairwise comparisons, and where it comes from.** Each entry is the fraction of comparisons on which a rule picks the video the annotators preferred. The model-identity rule never looks at the videos: it predicts whichever model wins more often overall, refitted on held-out prompts. The last column keeps only the pairs on which the two rules disagree.

	all pairs	where the two disagree
VBench subject consistency	0.824	0.423
model-identity rule (no video)	**0.860**	0.577

VBench's agreement with the annotators therefore comes mostly from pairs that the identity of the model already decides.

L.5 AGREEMENT WITH HUMAN RATINGS OF GENERATORS

VideoFeedback (He et al., 2024) rates 150 videos from each of 10 generators, and 150 clips collected from Sora, on a scale of 1 to 4 for temporal consistency, dynamic degree and visual quality. We use Spearman's ρ over the 10 generators between DED on the temporal axis (Appendix L.7), oriented so that higher is better, and the mean human temporal-consistency rating, supported if its 95% CI excludes zero. We score against two references, the MSR-VTT clips used in Appendix L.6 and the real clips distributed with VideoFeedback.

Table 12: **Agreement with human ratings of 10 generators from VideoFeedback.** Each entry is the Spearman correlation ρ, over the generators, between a score and the mean human rating in one category. All scores are oriented so that higher is better, so a positive ρ means the score orders the generators as people do. DED uses the temporal axis of Appendix L.7 with one of two real references.

	temporal consistency	dynamic degree	visual quality
DED, temporal axis, MSR-VTT reference	-0.67	$+0.47$	-0.62
DED, temporal axis, VideoFeedback reference	-0.53	$+0.53$	-0.50
VBench subject consistency	$+0.90$	-0.54	$+0.88$
VBench temporal flickering	$+0.93$	-0.52	$+0.81$
VBench motion smoothness	$+0.95$	-0.42	$+0.82$
VBench dynamic degree	-0.65	$+0.95$	-0.49

The test fails with either reference (Table 12). DED on the temporal axis orders these generators opposite to people's temporal-consistency ratings, significantly against MSR-VTT ($p = 0.03$) and not significantly against VideoFeedback's own clips (CI $[-0.88, +0.20]$), while VBench's per-video dimensions order them as people do. DED follows people's dynamic-degree ratings instead, at $+0.65$ ($p = 0.03$) when the Sora clips are included. Per video the picture is the same: within a generator, DED's typicality correlates with the human rating at -0.12 and VBench's subject consistency at $+0.40$.

The three generators people rate lowest, text2video-zero, zeroscope and lvdm, are also rated among the most dynamic (2.92 to 2.96 of 4). Their frames change a great deal, and the drift profile on which the temporal distance is built records how much and when a clip's representation moves, not whether the movement is coherent. DED can detect only what its distance measures, so this is a limitation of that distance rather than of the construction. A distance on point tracks or on the coherence of optical flow would be the natural replacement and needs no change to the method. Against MSR-VTT, every source, the Sora clips included, is off the reference support (95% to 100% of its videos), which is the failure of coverage described in Appendix L.8.

Two conclusions follow. On these generators, per-video consistency agrees with human judgment at the level of models, so the disagreement between DED and VBench across published models in Appendix L.7 should not be viewed as DED being closer to people. And the results of Appendix L.6, which do not depend on human ratings, are what the comparison with VBench rests on.

L.6 WHAT A MEAN OF PER-VIDEO SCORES CANNOT SEE

The substitutes above are built from real clips. Here we manipulate each generator's *own* pool instead, so that a manipulated pool differs from that model's clean pool only by the manipulation. We use the 20 generators of Appendix L.3, 5 clips for each of 20 prompts, and apply each manipulation to a fraction $q \in \{0, 0.25, 0.5, 0.75, 1\}$ of the pool, with the sets nested in q:

- **collapse**: for a fraction q of the prompts, every sample of the prompt is replaced by a copy of one of them, chosen uniformly.
- **reversal**: a fraction q of the clips is played backwards.
- **ping-pong**: the first half of a clip plays forward and then back to the start.
- **stutter**: every other frame is duplicated, $0, 1, 1, 2, 3, 3, \ldots$, keeping the length.

Every manipulation makes the pool worse, so detection is scored against the manipulation itself. A manipulation counts as detected for a model when the manipulated pool differs from the model's clean pool at $|z| > 2$ at some q. VBench's dimensions are recomputed per clip with its definitions, and motion is flow magnitude in place of the binary dynamic degree.

Table 13: **What a mean of per-video scores cannot see.** Each manipulation is applied to a growing fraction q of a generator's own clips, and each entry is the percentage of the 20 generators for which the manipulated pool differs from that generator's clean pool at $|z| > 2$ for some q. Every manipulation makes the pool worse, so a good score detects all four. A $+$ marks a VBench score that moves in the direction VBench rewards, monotonically in q, in every model (Spearman $\rho(q) = \pm 1.00$), so the manipulation is not only missed but rewarded. DED uses the temporal axis of Appendix L.7 with MSR-VTT as the reference. Every comparison is between two pools of the same generator, so the reference does not need to cover them.

	subj. consistency	flickering	motion	DED
collapse	15	0	20	**80**
reversal	0	0	0	**100**
ping-pong	15^{+}	0	0	**100**
stutter	35^{+}	100^{+}	30	**100**

Collapse is invisible to a per-video mean by construction: each kept video is still a real sample of its prompt, so the pool mean is unchanged in expectation over which sample is kept. We check this numerically: over 20 models the deviation of the pool mean from the clean mean, in units of its Monte Carlo standard error, has mean -0.03 and standard deviation 0.95 for subject consistency, as it should if the mean is unbiased. Flickering is exactly invariant to reversal, since $|x_t - x_{t+1}|$ is symmetric, and we confirm the change is 0 to floating point on every clip. Stutter makes a third of consecutive pairs identical, which lowers the frame change and raises the cosine similarity, so both VBench dimensions reward it. DED detects reversal, ping-pong and stutter in every model and collapse in 80% of them, because each changes the distribution of the model's clips relative to its clean pool, which is what a divergence measures.

L.7 The Three Axes

Appendices L.1 to L.3 examine what VBench's dimensions measure. This section applies our own construction instead, on three axes rather than one.

Each axis needs its own representation and its own distance:

- **temporal**: DTW on the drift profile of DINOv2 features, which records how much and when a clip's representation moves. It depends on frame order.
- **appearance**: Chamfer distance between per-frame CLIP embeddings treated as a *set*. It ignores frame order, so this axis is blind to exactly what the temporal axis sees.
- **flicker**: DTW on a profile of raw pixel differences between frames, with no learned representation.

The reference is the 600 real clips distributed with VideoFeedback (He et al., 2024), processed at the same fixed 8 frames per second as the generated clips and split into disjoint sets R (400 clips) and E (200). On the temporal axis it covers 16 of the 20 generators, with at most 3% of their clips off-support. CogVideoX-2B and 5B, EasyAnimate and Mira are not covered (80% to 92% off-support). MSR-VTT covers exactly these four and none of the others, so no available real corpus covers all 20 (Appendix L.8). Correlations below are therefore computed over the generators each axis covers, and VBench's scores are its official ones.

The axes rank different models best (Mochi on temporal, OpenSora 1.1 on appearance, Latte-1 on flicker) and are only partly related (Spearman -0.01 between temporal and appearance, $+0.48$ between temporal and flicker, -0.12 between appearance and flicker), so they are not three views of one quantity.

The temporal ranking disagrees with VBench. Over the 16 generators the reference covers, our temporal-axis DED correlates with VBench's subject consistency at $\rho = +0.52$ ($p = 0.04$, and

Table 14: **Two of the three distances fail negative type, but all three yield a divergence.** The first two numeric columns give the smallest eigenvalue of $-\frac{1}{2}HDH$ on R for d and for d^2 (Appendix M), where a negative value means that energy distance on that distance is not guaranteed to be valid. Chamfer on CLIP frame sets comes out at the eigensolver's noise floor for d and we do not count it as a failure, although d^2 fails. The last column counts the pairs of models that differ after Benjamini-Hochberg correction at $q = 0.05$.

axis	distance	d	d^2	separable pairs
temporal	DTW on DINOv2 drift	-9.9×10^{-1}	-1.3×10^{0}	166/190
appearance	Chamfer on CLIP frame sets	$\sim -10^{-16}$	-9.9×10^{-2}	150/190
flicker	DTW on pixel differences	-9.3×10^{-1}	-9.1×10^{-1}	181/190

$+0.39$ over all 20). DED is a divergence, so larger means further from real video: a positive correlation says the models VBench rates most consistent tend to be the ones DED places further from real data. VBench's three highest-scoring models, OpenSoraPlan 1.2, AnimateDiff-V1 and AnimateLCM, all covered, rank 7th, 19th and 17th of 20 under DED.

We do not claim our ordering is correct: on the ten generators VideoFeedback rates (Appendix L.5), people side with VBench's consistency, not with DED.

L.8 What the reference set has to satisfy

DED compares a generated pool against real data, so it inherits whatever the real data is, and the off-support rate says when the reference does not cover the candidates. Scored against MSR-VTT, every source in VideoFeedback, the clips collected from Sora included, has 95% to 100% of its videos off the reference support. Against the real clips distributed with VideoFeedback, the same sources have 0% to 5%, except text2video-zero at 11%, the generator people rate lowest (Appendix L.5). Against the first reference the off-support rate cannot separate the sources, and the divergence measures how each one's diffusion embedding lands relative to real data that none of them approaches. The check is available before any divergence is compared.

The ranking also depends on which real corpus the reference is drawn from. Kendall's τ between rankings of the 20 VBench generators is 0.77 to 0.80 when R and E are redrawn within one corpus and 0.46 between two corpora (Appendix I). Thus, the requirement is that the reference must cover the candidates in whatever the chosen distance is sensitive to, which the off-support rate measures directly, and a ranking must be reported with its choice of reference.

M Negative Type of Common Distances

We test negative type on a finite sample $x_1, \dots, x_n$ through its distance matrix $D_{ij} = d(x_i, x_j)$. With the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, the condition $\sum_{ij} c_i c_j D_{ij} \le 0$ for every c with $\sum_i c_i = 0$ is equivalent to $-\frac{1}{2}HDH \succeq 0$, since $Hc = c$ for such c and every Hv sums to zero. We report the smallest eigenvalue of $-\frac{1}{2}HDH$, computed for d and for d^2: a negative value certifies that the distance is not of negative type on that sample, while a value at the rounding floor of the eigensolver ($\sim 10^{-15}$ relative to the entries) is consistent with it. The values are in the units of the distance, so they are comparable only within a row.

Negative type is a strong constraint, and several familiar distances fail it. For example, $\mathbb{R}^n$ under the L_p norm for $p > 2$ fails the condition in dimensions $n \ge 3$ (Lyons, 2013), and on a curved Riemannian manifold the squared geodesic distance is not conditionally negative definite, so Gaussian kernels built on it are not positive definite (Feragen et al., 2015).

Squared Euclidean distance shows that being a metric, being of negative type and being of strong negative type are separate properties. It is not a metric, since it violates the triangle inequality, but it *is* of negative type: $\sum_{ij} c_i c_j \|x_i - x_j\|^2 = -2\|\sum_i c_i x_i\|^2 \le 0$. It is not of strong negative type, since $D_d^2(P, Q) = 2\|\mathbb{E}[X] - \mathbb{E}[Y]\|^2$, so it compares first moments and nothing else.

Table 15: **Negative type of common distances on real data.** Each entry is the smallest eigenvalue of $-\frac{1}{2}HDH$, computed from the distance matrix on n real samples, for the distance d and for its square d^2 (Appendix M). A negative value proves that the distance is not of negative type, so energy distance and distance-induced MMD built on it are not guaranteed to be valid. The values are in the units of each distance, so they can be compared only within a row. Exact optimal transport and SILog are at the rounding floor for d, and SILog, which is Hilbertian, is at the floor for d^2 as well.

modality	distance	d	d^2	n
point clouds	Chamfer	-8.1×10^{-1}	-3.5×10^{1}	200
depth	AbsRel	-9.5×10^{-2}	-4.5×10^{1}	300
video	DTW on flow	-3.4×10^{0}	-2.6×10^{2}	200
molecules	best-RMSD	-5.6×10^{-6}	-3.9×10^{-10}	200
ViT token sets	Hausdorff	-1.5×10^{2}	-4.8×10^{4}	250
ViT token sets	exact OT	-1.1×10^{-15}	-2.5×10^{1}	250
depth	SILog	-1.2×10^{-16}	-3.0×10^{-17}	200

A classical result, Schoenberg's theorem (Schoenberg, 1938), makes the connection to kernels precise: d is of negative type if and only if $\exp(-\lambda d)$ is a positive-definite kernel for every $\lambda > 0$.