WHAT DOES FRÉCHET DISTANCE MEASURE? A DIRECTIONAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Fréchet distance is a *de facto* standard for evaluating generative models across domains, appearing as FID for images and FVD for videos. It summarizes the discrepancy between generated and reference distributions in a single scalar, with lower values typically interpreted as better generation quality. However, this scalar view can obscure what drives the comparison. For example, in COCO dataset, increasing the number of diffusion sampling steps improves ImageReward scores yet worsens (increases) FID. Motivated by this mismatch, we seek to make the Fr'echet distance more interpretable by uncovering *where* the discrepancy lies. To this end, we introduce *directional Fréchet distance*, the expected squared projection of the optimal transport displacement onto a given direction. Across our image, video, and protein case studies, we find that a small number of interpretable directions account for much of the distance. We use these directions to explain the FID increase in terms of semantic concepts represented by CLIP embeddings, quantify FVD's bias toward per-frame appearance, and revisit the interpretation of Protein FID.

1 INTRODUCTION

Fréchet distance (Dowson & Landau, 1982) underlies *de facto* evaluation metrics for generative models in continuous domains: Fréchet Inception distance (FID) for images (Heusel et al., 2017), Fréchet video distance (FVD) for videos (Unterthiner et al., 2018), and Fréchet audio distance (FAD) for audio (Kilgour et al., 2018). A lower distance is read as a better generator, so the metric does not merely report progress but also *defines it*, steering architecture, sampling, and hyperparameter choices (Lu et al., 2025; Kim et al., 2025). This influential signal, however, remains a **scalar summary** of a comparison between two high-dimensional feature distributions.

This blind spot surfaces in recurring contradictions. For images, tuning a diffusion sampler configuration to minimize FID can drive per-sample quality to its worst (Lee et al., 2026); stronger image distortions can even improve (lower) the metric (Jayasumana et al., 2024); additional sampling steps can produce visibly sharper images (Choi et al., 2025), yet worsen FID on the COCO dataset (Lin et al., 2014). For videos, FVD favors per-frame appearance over motion, so a flickering clip can still score well (Ge et al., 2024).

Prior work has sought to address these issues by refining Fréchet metrics or introducing *scalar alternatives*, such as changing the embedding (Kynkäänniemi et al., 2022), adopting a kernel two-sample statistic (Bińkowski et al., 2018; Jayasumana et al., 2024), or reducing estimator bias (Chong & Forsyth, 2020). These changes can mitigate specific failures, but the resulting metrics still do not reveal *where* in feature space the distributions differ or *what* the discrepancy means.

Contribution. In this work, we take a complementary approach: making existing Fréchet metrics interpretable. We introduce **directional Fréchet distance**, which measures the contribution of a fixed feature direction to the Fréchet distance. These directional distances sum to the Fréchet distance over any orthonormal basis, and the top eigenvectors of the displacement second-moment matrix inform where the discrepancy concentrates. We further *affix* these directions to **interpretable concepts** using CLIP concept bridges (Kim et al., 2018; Bhalla et al., 2024) for images and structural-property concept directions for proteins.

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.

This gives **semantic meaning to directional distances**: a larger distance along a direction associated with a concept, e.g., composition simplicity or brightness, indicates that the metric attributes more of the discrepancy between generated and reference samples to that concept. With this lens, we revisit three established Fréchet metrics—FID, FVD, and Protein FID—to explain what drives previously reported discrepancies:

- Images: In some reference datasets such as COCO, additional sampling steps can increase ImageReward but also increase FID (Lee & Pak, 2026). We find that two-thirds of this increase lies along a single direction in the feature space associated with composition simplicity.
- Videos: In the controlled distortions of Ge et al. (2024), I3D-based FVD responds weakly to added flicker. We find that the flicker band carries less of the distance than the appearance band with I3D features, but more with VideoMAE-v2 features.
- Proteins: Protein FID was introduced to detect collapse onto a narrow range of easily foldable structures (Faltings et al., 2025), and its observed gap was partly attributed to missing beta sheets. We find that a single direction associated with structural regularity accounts for most of the gap, while the alpha-versus-beta direction accounts for little.

2 PRELIMINARIES

In Section 2.1, we review the Fréchet distance and the family of metrics built from it. In Section 2.2, we revisit prior observations across images and videos whose apparent contradictions stem from reading this single number as a measure of per-sample quality.

2.1 THE FRÉCHET DISTANCE AND ITS INSTANCES

Fréchet distance is widely used to evaluate generative models by comparing Gaussian approximations to generated and reference distributions in a fixed feature space. A prominent example is Fréchet Inception distance (FID) (Heusel et al., 2017), which uses InceptionV3 (Szegedy et al., 2016) features. The Fréchet distance is the squared 2-Wasserstein distance between the fitted Gaussians, which admits the closed form

$$\mathrm{FD} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right), \tag{1}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of the reference and generated feature distributions.

The same Fréchet distance can also be computed over other embeddings such as CLIP ViT-L/14 (Radford et al., 2021) and DINOv2 (Oquab et al., 2023). The same recipe extends the metric across domains: Video is evaluated with FVD (Unterthiner et al., 2018), the Fréchet distance over I3D embeddings (Carreira & Zisserman, 2017), and protein structure generation has recently adopted Protein FID (Faltings et al., 2025) over ESM3 structure-token embeddings (Hayes et al., 2025). Across all of these, a full comparison between two distributions is reduced to a single number, which reports *that* they differ but not *where*.

2.2 CONTRADICTORY OBSERVATIONS

A lower Fréchet distance is commonly interpreted as a *higher-quality* generator. This interpretation suggests that improvements in per-sample quality should be accompanied by a decrease in distance. In this section, we present two prior observations that challenge this.

To begin with, we first formally define the *per-sample quality*: for a chosen sample-level scoring function q, we denote the quality of generator G to be $\mathbb{E}_{x\sim G}[q(x)]$. Examples of such scoring functions include CLIP score (Hessel et al., 2021), ImageReward (Xu et al., 2023), and the numerical accuracy of each sample.

Images. Two standard knobs on Stable Diffusion 1.5 (SD1.5, Rombach et al., 2022) make this concrete (Figure 1). Raising the classifier-free guidance scale (Ho & Salimans, 2022) makes each sample match the prompt better and improves ImageReward, yet FID on the COCO reference

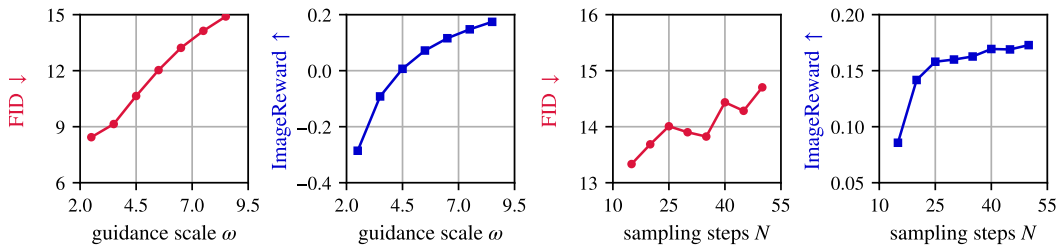


Figure 1: *Sign reversal* on SD1.5/COCO under two standard knobs. As the classifier-free guidance scale (left) or the number of sampling steps (right) increases, per-sample quality rises while FID worsens. A lower FID therefore does not track higher per-sample quality.

dataset (Lin et al., 2014) worsens monotonically (Lee et al., 2026, Table 8). Using more denoising steps reduces numerical error in each sample and also improves ImageReward (Lee & Pak, 2026), but FID can rise (Choi et al., 2025, Table 1) or converge (Lu et al., 2025, Section 7.2).

We call this contradictory phenomenon *sign reversal*: a change that raises per-sample quality raises the Fréchet distance instead of lowering it. Sign reversal makes FID an unreliable gate on generation quality. A sampling algorithm (Lu et al., 2025) or an inference accelerator (Kim et al., 2025) is accepted once it matches or lowers FID, read as evidence that quality is preserved. Under sign reversal, that inference is invalid, since the same low FID is equally consistent with per-sample quality having dropped, and the verdict on the method no longer follows.

Videos. FVD computes the Fréchet distance in the feature space of I3D (Carreira & Zisserman, 2017), a video model pretrained for action recognition. In practice, however, FVD favors per-frame appearance over motion (Ge et al., 2024): a clip whose individual frames are realistic can still score well when it flickers. To address this, Ge et al. (2024) propose replacing I3D features with self-supervised video embeddings such as VideoMAE-v2 (Wang et al., 2023b), improving sensitivity to motion while retaining the same Fréchet distance computation.

3 Directional Fréchet distance

The contradictory observations of Section 2.2 highlight a limitation of interpreting Fréchet distance (FD) as a measure of per-sample quality: FD aggregates discrepancies across feature directions into a single value, obscuring *which* of them drive it. We therefore introduce ***directional Fréchet distance***, the contribution of a given direction to FD, which recovers directional information that is encoded in the means and covariances used to compute FD but lost in the final number.

Directional Fréchet distance. To define directional Fréchet distance, we first express FD as an optimal transport cost. Assuming that two Gaussian distributions $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$, which are defined by the first and second moments of the reference and generated distributions, are both nondegenerate, the FD in equation 1 is defined by (Knott & Smith, 1984; Peyré & Cuturi, 2019)

$$\mathrm{FD} = \mathbb{E}[\|x - T(x)\|^2], \quad T(x) := \mu_g + A(x - \mu_r), \quad A = \Sigma_r^{-1/2} \left(\Sigma_r^{1/2} \Sigma_g \Sigma_r^{1/2}\right)^{1/2} \Sigma_r^{-1/2}.$$

The expectation is calculated over $x \sim \mathcal{N}(\mu_r, \Sigma_r)$. Here, one can show that T is the optimal transport map between $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$, i.e., it minimizes the expected squared norm among all measurable maps that transport $\mathcal{N}(\mu_r, \Sigma_r)$ to $\mathcal{N}(\mu_g, \Sigma_g)$:

$$T = \mathrm{argmin}_S \mathbb{E}_{x \sim \mathcal{N}(\mu_r, \Sigma_r)}[\|x - S(x)\|^2],$$

where the argmin is taken over all maps S satisfying $S(x) \sim \mathcal{N}(\mu_g, \Sigma_g)$ when $x \sim \mathcal{N}(\mu_r, \Sigma_r)$.

Now we define *directional Fréchet distance*. For a given unit direction w, the directional Fréchet distance $\mathrm{FD}(w)$ is computed as the squared distance *along* w under the optimal transport map T, i.e.,

$$\mathrm{FD}(w) := \mathbb{E}_{x \sim \mathcal{N}(\mu_r, \Sigma_r)}[(w^\top (x - T(x)))^2].$$

An alternative way to formulate $\mathrm{FD}(w)$ is to introduce the *displacement second-moment matrix* (Paty & Cuturi, 2019; Meng et al., 2020) $M = \mathbb{E}_{x \sim \mathcal{N}(\mu_r, \Sigma_r)} \left[(x - T(x))(x - T(x))^\top\right]$, resulting in $\mathrm{FD}(w) = w^\top M w$. Therefore, for any orthonormal basis $\{w_k\}$, directional Fréchet distances sum to the Fréchet distance, $\mathrm{FD} = \sum_k \mathrm{FD}(w_k)$ (Appendix B).

Interpretation. A large $\mathrm{FD}(w)$ indicates that displacement along w contributes substantially to the optimal transport cost between the generated and reference Gaussians. This decomposition thus provides a *diagnostic tool* for understanding what drives FD. For instance, when a direction is associated with an interpretable concept, such as brightness or composition simplicity, its directional FD quantifies how much of the discrepancy between the generated and reference distributions is attributed to that concept direction.

4 EXPERIMENTS

In this section, we use directional Fréchet distance to revisit three Fréchet metrics. We first explain why additional denoising steps raise ImageReward but also raise FID, and why the sign of this effect flips across references (Section 4.1). We then explain why VideoMAE-v2 makes FVD more sensitive to flicker than I3D (Section 4.2). Finally, we examine which structural directions drive the Protein FID gap (Faltings et al., 2025) (Section 4.3).

Each case study follows a common procedure. We take the top eigenvectors $e_1, e_2, \ldots$ of a displacement second-moment matrix M (Section 3), which identify the directions along which a distance concentrates. We then interpret these directions within each domain (Appendix C).

4.1 SIGN REVERSAL

We study sign reversal on images in the common setting of FID on the COCO reference, where additional denoising steps on Stable Diffusion 1.5 (SD1.5) raise ImageReward score but also raise FID (Figure 1). We build the explanation in two steps: (1) what the top directions mean, (2) then which of them carries the change in FID and why its sign flips across references.

We first analyze a single comparison, SD1.5 at 50 steps against COCO, to find the directions FID concentrates on and name them with concept dictionaries (Section 4.1.1). We then sweep the denoising steps and follow which of these directions carries the change in FID, first on COCO and then across six references, to explain what sets its sign (Section 4.1.2).

4.1.1 NAMING THE TOP DIRECTIONS

We evaluate SD1.5 at 50 denoising steps against the COCO reference in Inception feature space, where FID is 14.6, and ask three questions of this single comparison: where in feature space the distance concentrates, what those directions mean, and how the two distributions differ along them.

Where: FID concentrates on a few directions. Ordering the eigenvectors $e_1, e_2, \ldots$ of M by their directional FD ranks the feature directions by the share of FID they carry, and the spectrum is steep: e_1 carries 24.0% of FID, e_2 carries 3.8%, and e_3 carries 2.5%, so three eigenvectors account for 30.3% of a distance computed across 2048 feature dimensions.

What: CLIP as a bridge. A substantial fraction of FID thus concentrates on three directions, but these are directions in Inception feature space and have no semantics on their own, so naming them is what turns the decomposition into a statement about how the reference and generated sets differ. We attach meaning with CLIP as a bridge between text and the Inception features FID is computed in. We write a CLIP dictionary of 24 photographic concepts, each a pair of opposing prompt sets, e.g. a single clean subject against a cluttered scene, and embed both poles with the CLIP text encoder.

For each concept, we rank the reference images by CLIP similarity to its text direction and take the top and bottom quantiles as the aligned and anti-aligned groups, and the difference of their mean Inception features is the concept direction w_c, a supervised direction in the same space as the eigenvectors, similar to the concept-activation construction of Kim et al. (2018). Given the concept directions, we fit each eigenvector as a sparse signed combination of them with a Lasso and take the surviving concepts as the direction's name, as in Bhalla et al. (2024). Table 1 lists the result: e_1 reconstructs with cosine similarity 0.74 and is dominated by composition simplicity, so we name it after that concept; it runs from cluttered multi-object scenes toward simple single-subject images. e_2 reconstructs more weakly at 0.49 as a mixture of brightness, weather, and indoor-outdoor content.

What: ImageNet classes name what CLIP concepts miss. The CLIP dictionary does not reach e_3, which reconstructs to cosine similarity 0.17 with no dominant concept, so we turn to a sec-

Table 1: Sparse concept fit of the top eigenvectors of M over the two concept dictionaries, for SD1.5, 50 steps, and COCO. Share is $\mathrm{FD}(e_i)/\mathrm{FD}$, the weights are the largest surviving Lasso coefficients, and cos is the reconstruction cosine.

	share	dictionary	surviving concepts (weight)	cos
e_1	24.0%	CLIP	composition simplicity (0.55), depth of field (0.12), furniture (0.08)	0.74
e_2	3.8%	CLIP	brightness (-0.15), weather (-0.12), crowd density (0.08), indoor/outdoor (0.08)	0.49
e_3	2.5%	ImageNet	Dungeness crab (-0.09), mashed potato (-0.06), burrito (-0.05)	0.40

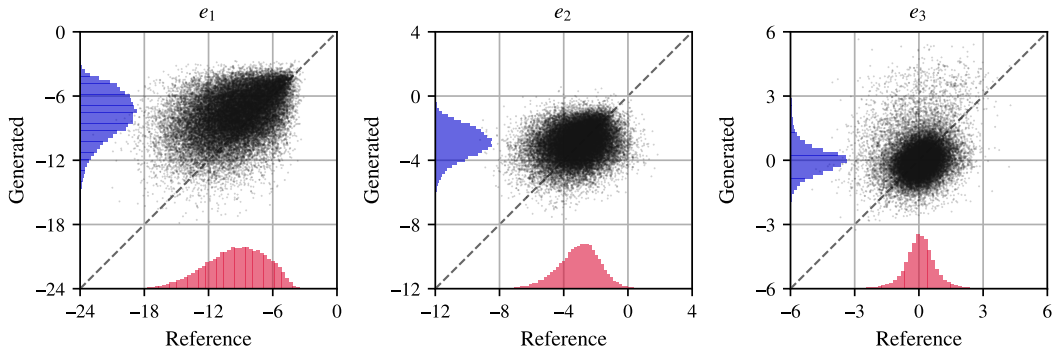


Figure 2: Reference and generated features projected onto the top three eigenvectors of M, using SD1.5, 50 steps, and COCO. Each point pairs a caption's reference and generated projection. The reference (red) and generated (blue) marginal distributions are drawn inside, and the dashed line is $y = x$. For e_1, the points are slightly above $y = x$, indicating that the generated set is shifted toward $+e_1$. For e_2, the points concentrate vertically, demonstrating that the generated set has less variance, and vice versa for e_3.

ond dictionary built without prompts. The Inception features map to the 1008 ImageNet logits through one affine layer, so the unit-normalized rows of that layer form a data-free concept dictionary (Kynkäänniemi et al., 2022). A sparse fit over these class rows names e_3 as a consistent food cluster (Dungeness crab, mashed potato, burrito, ice cream), so together the two dictionaries name the top three eigenvectors of M.

How: a shift in mean or a change in spread. Projecting the two sets onto each e_i (Figure 2) tells whether they differ in mean or in spread along it. On e_1, the generated mean sits toward the simple-composition end (Cohen's $d = 0.69$), so the generated set is shifted toward simpler, more single-object images than COCO. On e_2 and e_3, the means coincide and only the spread differs: the generated set has 0.56 times the reference variance along e_2, holding a narrower range of lighting and indoor-outdoor content, and 1.16 times along e_3, spreading more widely over food content.

Finding. The direction that carries the most directional FD, e_1, is dominated by composition simplicity, and the generated set differs from COCO by a shift along it, toward simpler, single-object images.

Robustness check: DINOv2. To test whether this finding holds across feature spaces, we repeat the decomposition in DINOv2 feature space (Oquab et al., 2023). The dominant direction has a similar interpretation: e_1 captures composition simplicity, accounts for 14.0% of the distance, and reconstructs at a cosine similarity of 0.41. The generated set is again shifted relative to COCO toward simpler compositions (Cohen's $d = 2.2$). Thus, the interpretation of the dominant direction is consistent across embeddings.

4.1.2 Which directions carry the reversal

We next use this decomposition to identify which directions account for sign reversal. Using SD1.5 and COCO again, we take eight generated sets using $\{15, 20, \ldots, 50\}$ steps.

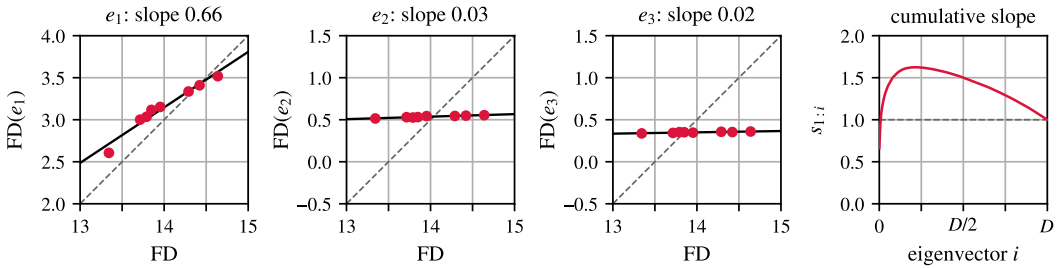


Figure 3: *Left*: directional FD of the top three fixed eigenvectors against total FID across the eight-step COCO sweep, in the frozen 50-step basis. Each point is one step setting, the solid line is the least-squares fit, and its slope is the fraction of the FID change that eigenvector carries, and the dashed line is the slope $= 1$ an eigenvector carrying the entire change would follow. The top eigenvector e_1 carries most of it (slope 0.66, $R^2 = 0.96$), while e_2 and e_3 stay nearly flat (slopes 0.03 and 0.02). *Right*: cumulative slope $s_{1:i}$, the summed slopes of the top i eigenvectors, which rises above the whole change (dashed) over the head and falls back to it over the tail (the eigenvectors before and after the curve's peak).

To track the same directions across these settings, we fix the eigenbasis of the 50-step M and regress each direction's $\mathrm{FD}(e_i)$ against the total FID; since the directional FDs over the basis sum to FD, the slopes sum to one, and each slope is the fraction of the FD change the direction carries (Figure 3).

One eigenvector accounts for two thirds of the increase. The top eigenvector e_1 has slope 0.66 ($R^2 = 0.96$), while e_2 and e_3 stay nearly flat at slopes 0.03 and 0.02, so two thirds of the FID increase lives on e_1 alone. FID thus rises with steps mainly because the generated mean drifts further toward simpler, single-object images and away from COCO, which ties most of the sign reversal of Section 2.2 to one named direction (Appendix D).

Rise over the head, fall over the tail. The slope regression says how much each eigenvector carries but not where the change accumulates along the spectrum. To see that, we sum the slopes of the top i eigenvectors into the cumulative slope $s_{1:i}$ and plot it against i (Figure 3, right). Since the directional FDs of all D eigenvectors sum to FID, $s_{1:D} = 1$. The curve rises above one, peaking at 1.62 near $i = 440$, and then falls back to one; we call the eigenvectors up to this peak the *head* and the rest the *tail*. The rise opens with the jump of e_1, along which the generated mean drifts further toward simpler images as steps increase (Cohen's d grows from $+0.58$ to $+0.69$), and continues with small gains over the other head eigenvectors. Over the tail the curve falls, as more steps contract the generated variance there by about 1.5% from 15 to 50 steps, lowering directional FD by small amounts across the tail eigenvectors. On COCO the rise outweighs the fall, so FID rises overall (sign reversal). Appendix E repeats this sweep on six references.

> **Finding.** Two thirds of the step-driven FID increase on COCO lies on a single eigenvector, e_1, named after composition simplicity, along which the generated mean drifts toward simpler images. Over the whole spectrum, the rise over the head outweighs a broad fall over the tail.

4.2 Video content bias

Fréchet video distance (FVD, Unterthiner et al., 2018) is the Fréchet distance over the features of I3D (Carreira & Zisserman, 2017), which is an action-recognition model. Ge et al. (2024) expose the *content bias* of FVD (Section 2.2) by decoupling per-frame appearance from motion. In particular, they create two sets of degraded real videos: one with the same distortion across frames and another with a distortion of matched severity resampled each frame. They find that I3D-based FVD barely penalizes the resulting flicker.

To address this bias, they replace I3D with VideoMAE-v2 (Wang et al., 2023b), which makes FVD more sensitive to flicker. We use directional FD to explain this difference: in I3D feature space, appearance-related directions account for more of the spatio-temporally distorted set's distance from real videos than flicker-related directions, whereas in VideoMAE-v2 feature space, the ordering reverses.

Table 2: Share of directional FD on the appearance and flicker bands (the top eigenvectors of $M(R, S)$ and $M(S, T)$), at the top $k = 1$ and $k = 10$ eigenvectors. The flicker-set panel evaluates $M(R, T)$ at severity 3, with bands averaged over the other four severities. The generator panel evaluates ModelScope's generated set, zero-shot on the UCF-101 class names, on bands averaged over all five severities. The two bands overlap, so the shares of a row can sum past one.

	flicker set, $M(R, T)$				generator, $M(R, G)$			
	$k = 1$		$k = 10$		$k = 1$		$k = 10$	
embedding, family	appearance	flicker	appearance	flicker	appearance	flicker	appearance	flicker
I3D, motion blur	0.325	0.023	0.663	0.273	0.022	0.005	0.288	0.150
I3D, elastic	0.372	0.211	0.648	0.495	0.027	0.012	0.288	0.107
VideoMAE, motion blur	0.061	0.470	0.181	0.612	0.013	0.019	0.095	0.096
VideoMAE, elastic	0.163	0.389	0.282	0.553	0.012	0.009	0.148	0.084

Decoupled distortions as probes. We obtain these directions from the two degraded sets of Ge et al. (2024), which we rebuild on a reference set R of 2048 UCF-101 clips of 16 frames (Soomro et al., 2012). The distortions are elastic transformation and motion blur, each at the five severities of ImageNet-C (Hendrycks & Dietterich, 2019). The *appearance set* S is their spatially distorted set, which applies one sampled distortion to every frame of a clip. The *flicker set* T is their spatio-temporally distorted set, which resamples the distortion at every frame. The two sets share the distribution of per-frame distortions, so T differs from S only by the flicker.

Appearance and flicker displacements. We write $M(P, Q)$ for the displacement second-moment matrix between two sets P and Q. The *appearance displacement* $M(R, S)$ reflects the distortion itself. The *flicker displacement* $M(S, T)$ reflects the flicker added on top of it. We average each displacement over severities and take its top eigenvectors as the *appearance band* and the *flicker band*. The cumulative share $\sum_{i \le k} \mathrm{FD}(e_i)/\mathrm{FD}$ is the fraction of a set's FD that the top k eigenvectors e_i of a band carry. The two bands are computed independently and can overlap. A share therefore measures how much of a distance aligns with a band, not an exclusive allocation between the bands (Appendix F).

4.2.1 Which band carries the flicker set's distance

We evaluate the displacement $M(R, T)$ between the reference and the flicker set on both bands. The share on the appearance band shows how much of this distance aligns with the distortion itself. The share on the flicker band shows how much aligns with the flicker.

The two embeddings put the flicker set's distance on opposite bands. Table 2 gives both shares at severity 3, with bands averaged over the other four severities. I3D puts far more of this distance on the appearance band than on the flicker band: with motion blur, 0.33 against 0.02 on the top eigenvector and 0.66 against 0.27 at $k = 10$. I3D thus sees the flicker set mostly through the distortion it shares with the appearance set. VideoMAE reverses the ordering, with 0.06 against 0.47 on the top eigenvector and 0.18 against 0.61 at $k = 10$. Both distortion families show this flip, and motion blur shows it more clearly. The flip with motion blur also holds when either band is orthogonalized against the other (Appendix F).

The band share holds where the scalar ratio fades. Ge et al. (2024) measure the bias with the ratio of the flicker set's FVD to the appearance set's. This ratio shrinks toward 1 as severity grows, because the appearance set's FVD grows faster than what the flicker adds. I3D's ratio with motion blur falls from 1.41 at the mildest severity to 1.07 at the harshest. The band share is a different quantity, and it keeps the two embeddings apart at every severity. VideoMAE's share on the top motion blur flicker eigenvector stays above 0.23, while I3D's stays under 0.06 (Appendix F).

Finding. I3D puts more of the flicker set's distance on the appearance band, the directions the distortion already moves, than on the flicker band. VideoMAE does the reverse. The content bias of FVD is this small flicker-band share under I3D.

4.2.2 A REAL GENERATOR ON THE BANDS

The bands so far come from controlled distortions. We now ask whether the content bias found on the flicker set (Section 4.2.1) also appears in a real generator's discrepancy from the reference. We sample a set G of 2048 clips with ModelScope text-to-video (Wang et al., 2023a), zero-shot on the UCF-101 class names. We then evaluate its displacement $M(R, G)$ on the appearance and flicker bands under both embeddings.

VideoMAE shifts the generated distance toward the flicker band. I3D puts about twice as much of $M(R, G)$ on the appearance band as on the flicker band, with 0.29 against 0.15 on the motion blur bands at $k = 10$ (Table 2). This is the same ordering as on the flicker set. VideoMAE splits the same distance evenly, with 0.095 against 0.096. The ratio of the flicker share to the appearance share thus rises from 0.52 under I3D to 1.01 under VideoMAE. The elastic bands show the same shift, from 0.37 to 0.57, although appearance still leads there (Appendix F).

> **Finding.** VideoMAE shifts a real generator's distance toward the flicker band. The two motion blur bands carry it evenly under VideoMAE, while I3D puts twice as much on the appearance band. A scalar ratio cannot report this shift upon an arbitrary generated set.

4.3 PROTEIN FID ATTRIBUTION

Designability is the most widely used per-sample metric for protein backbones: An inverse-folding model designs a sequence for the backbone, which is then refolded by a structure predictor. The backbone is considered designable if the predicted structure matches it within a self-consistency RMSD (scRMSD) of 2 Å (Trippe et al., 2022; Watson et al., 2023).

High designability, however, can be achieved by generating mostly alpha-helical backbones: alpha helices rely largely on local hydrogen bonds, whereas beta sheets require long-range coordination. Optimizing for designability can therefore favor a narrow range of easily foldable, helix-rich structures while losing beta sheets from the generated distribution.

To detect such collapse, Faltings et al. (2025) introduce Protein FID, the Fréchet distance between generated and reference structures embedded with the ESM3 structure encoder (Hayes et al., 2025). They validate it through its monotone response to structural perturbations and its recovery of the FoldSeek and CATH hierarchies.

The authors partly attribute the Protein FID gap to missing beta sheets, based on a sample-level influence analysis that identifies large alpha-helical bundles as major contributors. We use directional FD to examine which structural directions drive the gap: the alpha-versus-beta contrast accounts for only a small share, whereas a single direction associated with structural regularity accounts for the largest share.

Reference and embedding. Protein FID mean-pools the per-residue ESM3 embeddings to 1536 dimensions. It then applies a PCA to 32 dimensions, fit jointly over the reference and generated embeddings, because the raw covariances are nearly degenerate (Faltings et al., 2025). We analyze this 32-dimensional space and recompute the optimal transport in it. The reference is the released embeddings of curated PDB structures. The raw space gives the same picture (Appendix G).

Generations and concept directions. We generate unconditional backbones with La-Proteina (Geffner et al., 2026), 467 structures per setting, following the Protein FID protocol. To interpret the eigenvectors of M, we build a concept dictionary in the same space. We define the alpha-versus-beta direction using the mainly-alpha and mainly-beta CATH classes (Sillitoe et al., 2021), and additional concepts by splitting ten whole-structure geometric properties into terciles. Each concept direction is a difference between group means of CATH embeddings, projected onto the same PCA basis. We then fit each eigenvector with a Lasso over this dictionary, as in Section 4.1.1.

4.3.1 NAMING THE DOMINANT DIRECTION

We analyze La-Proteina at 400 steps against the reference and ask the three questions of Section 4.1.1: where, what, and how.

Table 3: Sparse concept fit of the top eigenvectors of M over the geometry concept dictionary, and how the two sets differ along each, for La-Proteina at 400 steps against the ESM3 reference, in the 32-dimensional PCA space of Protein FID. Share is $\mathrm{FD}(e_i)/\mathrm{FD}$, the weights are Lasso coefficients, cos is the reconstruction cosine, d is the Cohen's d of the reference and generated projections onto e_i, and var is their variance ratio. We show the top surviving concepts here. Refer to Appendix G for details.

	share	surviving concepts (weight)	cos	d	var
e_1	66%	coil (-0.97), contact order ($+0.53$), length ($+0.16$), globularity (-0.16)	0.72	$+3.2$	0.55
e_2	14%	backbone-bond deviation (-0.58), clash ($+0.50$)	0.92	$+0.0$	0.05
e_3	4%	contact order (-0.90), clash ($+0.31$)	0.83	$+0.1$	0.16

Where: the distance concentrates on one direction. The spectrum of M is steep: e_1 carries 66% of Protein FID, e_2 carries 14%, and e_3 carries 4%. The top direction is not an arbitrary shift of the generator away from the reference. The reference itself varies along e_1 more than along 94% of its principal directions, so e_1 is a significant direction of the reference data.

What: structural regularity. Table 3 fits the top eigenvectors over the concept dictionary. e_1 reconstructs at cosine 0.72 and is dominated by a coil deficit. The concept directions most aligned with e_1 on their own are coil (-0.46), helix ($+0.32$), and clash (-0.29). The generated end of e_1 is thus the regular, idealized one, and we name e_1 the *structural-regularity direction*. The lower eigenvectors read as backbone-bond uniformity and clash (e_2) and as the contact-order residual (e_3).

What: alpha versus beta is a minor part. The alpha-versus-beta direction carries only 5.5% of Protein FID and aligns weakly with the mean shift (cosine 0.24). It stays minor even though the generated set is extreme in composition: 71% helix and 5% strand, with 58% of samples containing no beta sheet at all. Protein FID charges these helix-rich samples mainly along the structural-regularity direction, and only a little along the alpha-versus-beta direction.

How: a shift in mean or a change in spread. Projecting the two sets onto each e_i tells whether they differ in mean or in spread along it. The generated mean sits toward the regular end of e_1 (Cohen's $d = 3.2$), a shift onto over-idealized backbones. The means coincide along e_2 and e_3, and only the spread differs. The generated set holds only 0.05 and 0.16 times the reference variance there. Its samples are nearly identical in backbone uniformity and contact order, where real proteins vary.

Finding. The dominant Protein FID direction, e_1, is the structural-regularity direction. It carries two thirds of Protein FID, and the generated set is shifted along it toward low-coil, compact, idealized backbones. The alpha-versus-beta direction carries only 5.5%, so Protein FID charges over-regularity far more than the missing beta sheets.

5 Conclusion

In this work, we introduced directional Fréchet distance, decomposing Fréchet distance along an orthonormal basis of feature directions. Across images, videos, and proteins, this viewpoint facilitates explaining the sign reversal of FID and the content bias of FVD, and revisits a prior interpretation of Protein FID.

Outlook. More broadly, we hope this work prompts a rethinking of what scalar evaluation can and cannot tell us. A scalar score is useful *per se*–it makes models easy to compare, spurring competition and progress in generative modeling. Nevertheless, it does not reveal which aspects of a distribution a model fails to capture. Directional FD opens the door to more interpretable evaluation of generative models.

AI USE STATEMENT

In this work, we used generative AI tools for a number of tasks. Among the tasks with required disclosure, we used AI tools to help develop the conceptual framework of directional Fréchet distance, to formulate and prove the mathematical claims of Section 3 and to draft the derivations that establish them, to propose and refine our hypotheses about the sign reversal mechanism, to design and give feedback on the experimental methodology, to implement the methods and the evaluation code, to clean and reformat the datasets used for evaluation, to support the qualitative and thematic analysis that names the directions, and to interpret the results. We did not use generative AI tools to generate synthetic datasets: every generated sample analyzed in this work is produced by the diffusion models under study. Assisting with translation is not applicable to this work. Additionally, among the tasks with recommended disclosure, we used generative AI tools to create and edit software code, to create and modify scientific figures, to draft and to edit parts of the manuscript for readability, to summarize and analyze existing literature, to suggest the structure of the paper, to propose candidate titles and keywords, to brainstorm, and to suggest experimental parameters.

We have reviewed all AI-assisted work: the authors verified and tested all AI-written code for correctness, checked every mathematical claim and derivation, and read and revised all AI-drafted text against the underlying experiments and the cited sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes an existing evaluation metric and introduces no new generative model or dataset. All experiments use publicly available datasets and pretrained models, and none involves human subjects.

REPRODUCIBILITY STATEMENT

We will release our code as supplementary material. All samples are generated and all features are computed with deterministic CUDA algorithms and fixed per-sample seeds (`experiments/generate.py`), so that rerunning the code on the same hardware reproduces them exactly. Appendix C describes the datasets and their preprocessing.

REFERENCES

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

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

Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

Jinyoung Choi, Junoh Kang, and Bohyung Han. Enhanced diffusion sampling via extrapolation with multiple ode solutions. In *International Conference on Learning Representations*, volume 2025, pp. 98852–98876, 2025.

Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078. IEEE, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

David C Dowson and B V Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Felix Faltings, Hannes Stark, Tommi Jaakkola, and Regina Barzilay. Protein fid: Improved evaluation of protein structure generative models. *arXiv preprint arXiv:2505.08041*, 2025.

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.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 67350–67395, 2026.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 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.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Daeun Kim, Jinwoo Hwang, Changhun Oh, and Jongse Park. Mixdit: Accelerating image diffusion transformer inference with mixed-precision mx quantization. *IEEE Computer Architecture Letters*, 24(1):141–144, 2025.

Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Yunghee Lee and Byeonghyun Pak. Rethinking fid through the geometry of the reference dataset. *arXiv preprint arXiv:2605.29335*, 2026.

Yunghee Lee, Byeonghyun Pak, Junwha Hong, and Hoseong Kim. Tortoise and hare guidance: Accelerating diffusion model inference with multirate integration. *Advances in Neural Information Processing Systems*, 38:2871–2898, 2026.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, and Wenxuan Zhong. Sufficient dimension reduction for classification using principal optimal transport direction. *Advances in neural information processing systems*, 33:4015–4028, 2020.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

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.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14277–14286, 2023.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11400–11410. IEEE, 2022.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pp. 5072–5081. PMLR, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature methods*, 22(10):2107–2117, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560. IEEE, 2023b.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

A Related Work

Critiques of FID. A line of work critiques the estimator and the space it lives in. Inception embeddings of standard datasets fail multivariate normality tests, violating the Gaussian assumption (Jayasumana et al., 2024). FID is a biased estimator whose bias depends on the model being evaluated, not just on the sample count (Chong & Forsyth, 2020). Differences in image resizing also shift it by several points on identical data (Parmar et al., 2022). The representation matters as much as the estimator. The second-last layer of InceptionV3 is one affine transform from the ImageNet logits, so FID is sensitive to ImageNet class information that need not reflect perceptual quality (Kynkäänniemi et al., 2022). The semantics an embedding encodes also depend strongly on its training (Stein et al., 2023). Surveys of evaluation practice add that the choice of metric is rarely motivated and that automatic metrics often disagree with human judgment (Borji, 2022; Otani et al., 2023).

Swapping one scalar metric for another. Each pathology has typically been mitigated by replacing one scalar metric with another. Examples include a cleaner embedding (Kynkäänniemi et al., 2022), a different kernel such as the MMD tests of KID (Bińkowski et al., 2018) and CMMD (Jayasumana et al., 2024), a debiased estimator (Chong & Forsyth, 2020), and a content-debiased FVD that traces the video bias to the embedding (Ge et al., 2024). Protein FID (Faltings et al., 2025) is the same move in a new domain, swapping the per-sample designability score for a distributional metric. These works support their diagnoses with further analyses. The metric each proposes, however, still reports a single number, which does not show where in feature space the discrepancy lives.

Reporting more than one number. The closest attempts to report more than one number measure a generator along two fixed criteria. Precision and recall separate sample fidelity from distributional coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), and density and coverage make those two criteria more robust to outliers (Naeem et al., 2020). These measures are informative but preset. They score a generator on fidelity and diversity, not along data-driven or semantically named directions.

The displacement second-moment matrix. The matrix M whose directions we name is not new to optimal transport. Paty & Cuturi (2019) define the second moment $V_\pi = \int (x-y)(x-y)^\top d\pi(x,y)$ of the displacements under a coupling π and note that its trace is the transport cost. They build subspace-robust Wasserstein distances from its top eigenvalues, optimizing the coupling and the subspace jointly. Meng et al. (2020) use the eigenvectors of the same second moment as a supervised projection for dimension reduction. Both use the matrix to define a new distance or a new projection. We instead keep the coupling fixed at the Gaussian optimal map behind an existing metric. $\operatorname{Tr} M = \mathrm{FD}$ then makes each direction an exact additive share of the Fréchet distance that an evaluation already reports.

Interpreting feature directions. A separate literature interprets the directions of a feature space rather than the metric. Concept activation vectors probe a representation along human-named concept directions (Kim et al., 2018), SpLiCE decomposes CLIP embeddings into a sparse sum of interpretable concepts (Bhalla et al., 2024), and sparse autoencoders recover interpretable features from protein language models (Simon & Zou, 2025). We reuse this machinery, but to name the directions a Fréchet metric charges for rather than to interpret an embedding in isolation.

B Methodological details

Recovering the closed form. Write $\delta = \mu_r - \mu_g$ and $C = (\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2})^{1/2}$. The map T of Section 3 then has $A = \Sigma_r^{-1/2} C \Sigma_r^{-1/2}$, a symmetric matrix with $A\Sigma_r A = \Sigma_g$ and $\operatorname{Tr}(A\Sigma_r) = \operatorname{Tr} C$. The displacement is $x - T(x) = \delta + (I - A)(x - \mu_r)$. The term $x - \mu_r$ has zero mean and covariance Σ_r, so the cross terms vanish in expectation and

$$M = \delta\delta^\top + (I-A)\Sigma_r(I-A), \qquad \mathrm{FD}(w) = (w^\top\delta)^2 + w^\top(I-A)\Sigma_r(I-A)\,w. \tag{2}$$

Taking the trace recovers the closed form equation 1,

$$\operatorname{Tr} M = \|\delta\|^2 + \operatorname{Tr}\Sigma_r - 2\operatorname{Tr}(A\Sigma_r) + \operatorname{Tr}(A\Sigma_r A) = \|\mu_r - \mu_g\|^2 + \operatorname{Tr}\left(\Sigma_r + \Sigma_g - 2C\right).$$

We call the first term of $\mathrm{FD}(w)$ in equation 2 its *mean part*, the mean gap read along w. We call the second term its *covariance part*, what the map pays along w to carry the reference spread onto the generated one. Summed over an orthonormal basis, the two parts give the two terms of the closed form. The mean part of M has rank one, so a dominant mean gap draws the top eigenvector e_1 toward the mean-difference direction $\delta/\|\delta\|$. The covariance part then decides the eigenvectors that follow.

Directional FD is not the FD of a projection. The pair $(w^\top x, w^\top T(x))$ with $x \sim \mathcal{N}(\mu_r, \Sigma_r)$ is one coupling of the two projected Gaussians. Its cost $\mathrm{FD}(w)$ is therefore at least their optimal cost, the Fréchet distance $(w^\top\delta)^2 + (\sqrt{w^\top\Sigma_r w} - \sqrt{w^\top\Sigma_g w})^2$ between the projections. The inequality can be strict, because the optimal coupling of the full distributions need not couple the projections optimally. For example, two covariances may differ only in how their coordinates correlate. Their projections onto each coordinate then coincide, but the map still pays along every coordinate to rotate the correlation. The projected means and variances, which we report as Cohen's d and the variance ratio, thus describe how the two sets differ along w. The same bound holds for a subspace. The directional FD of a subspace with orthogonal projector P is $\mathrm{Tr}(PM)$, the sum of $\mathrm{FD}(w)$ over an orthonormal basis of the subspace. Protein FID is the Fréchet distance between the projections onto a 32-dimensional PCA subspace. It is therefore not the directional FD of that subspace in the raw space, but a lower bound on it.

Singular covariances. Section 3 assumes both Gaussians nondegenerate, but the map needs only a nondegenerate Σ_r. A is then symmetric positive semidefinite for any positive semidefinite Σ_g, with $A\Sigma_r A = \Sigma_g$. T thus carries $\mathcal{N}(\mu_r, \Sigma_r)$ to $\mathcal{N}(\mu_g, \Sigma_g)$. T is also the gradient of a convex function, so by Brenier's theorem (Brenier, 1991) it is the unique optimal map from the absolutely continuous source. This covers the raw 1536-dimensional protein analysis of Appendix G. There the covariance of 467 generated structures has rank at most 466, while that of the 4943 reference structures is nondegenerate. Both covariances are nondegenerate in the 32-dimensional PCA space of Protein FID and in every image and video comparison.

Numerical computation. We compute in float64 and take every matrix square root from an eigendecomposition, which is real for any positive semidefinite input. Most implementations of FD instead apply `scipy.linalg.sqrtm`, a Schur decomposition, to the non-symmetric product of the two covariances. This returns spurious imaginary parts when the product is singular, although the true square root is real. The reference implementation of Protein FID (Faltings et al., 2025) is one of them and rejects the result once the imaginary part exceeds a fixed tolerance. It therefore fails on a generated set of 32 structures, whose covariance is singular in the 32-dimensional PCA space, and succeeds on 40. Directional FD also needs the inverse square root in A. We avoid it by assembling the covariance part of M as $\Sigma_r + \Sigma_g - K - K^\top$, where $K = \mathbb{E}[(x-\mu_r)(T(x)-\mu_g)^\top] = \Sigma_r A$ is the cross-covariance under the optimal coupling. We obtain $K = \Sigma_r^{1/2} W^\top \Sigma_g^{1/2}$ from the polar decomposition $\Sigma_g^{1/2}\Sigma_r^{1/2} = WP$ into an orthogonal W and a positive semidefinite P. This polar decomposition requires no inversion and exists for singular covariances too.

C Experimental details

Shared analysis. Each eigenvector e_i of M is oriented toward the generated mean, $e_i^\top(\mu_g - \mu_r) \geq 0$, so $+e_i$ is the generated end. Cohen's d along a direction is the difference between the means of the generated and reference projections, divided by their pooled standard deviation. The variance ratio is the variance of the generated projections over that of the reference projections. A concept direction w_c is the unit-normalized difference between the mean features of two groups of real samples, computed in the space that defines the metric. The groups are reference images ranked by CLIP image-text similarity for images, and CATH domains split by class or by a geometric property for proteins. The sparse concept fit regresses e_i on the concept directions of one dictionary with a Lasso without intercept, $\min_\beta \frac{1}{2D}\|e_i - W\beta\|^2 + \alpha\|\beta\|_1$, where the columns of W are the concept directions in D dimensions. It reports the surviving signed weights and the cosine between e_i and its reconstruction $W\beta$. A weight is positive when its concept grows toward $+e_i$. We set α to leave a handful of concepts: 5×10^{-5} for the CLIP dictionary, 7×10^{-5} for the ImageNet-class dictionary, and 10^{-3} for the protein geometry dictionary (10^{-5} in the raw space of Appendix G). Correlated

concepts can exchange weight in such a fit. A weight therefore need not match the alignment of its concept direction alone with e_i, and we name a direction by its dominant concepts rather than by every surviving weight.

Images. We generate the sets with SD1.5 (Rombach et al., 2022) using the DDIM sampler (Song et al., 2020) at classifier-free guidance scale 7.5 and 512×512 resolution, at $15, 20, \ldots, 50$ steps, with one image per reference caption. The COCO reference is the $30{,}000$ captioned images of the 2014 validation split. The Inception features are the 2048-dimensional pool features of clean-fid (Parmar et al., 2022). The CLIP dictionary uses CLIP ViT-L/14 at 336 pixels (Radford et al., 2021). It embeds each pole of a concept as the mean CLIP text embedding of its prompt set. The two groups are the top and bottom 10% of reference images by the cosine similarity of their CLIP image embedding to the difference of the two pole embeddings. The 24 concepts are composition simplicity, number of objects, crowd density, people, indoor/outdoor, nature versus urban, animals, food, vehicles, furniture, electronics, sports equipment, close-up versus wide shot, colorfulness, warm versus cool tone, brightness, water and sky, text and signs, staged versus candid, centered composition, depth of field, weather, daytime, and portrait versus scene. The ImageNet-class dictionary is the 1008 unit-normalized rows of the final Inception layer, which maps the pool features to the class logits. The DINOv2 check uses the ViT-L/14 model (Oquab et al., 2023) with the same dictionary and penalty.

Video. The reference is 2048 UCF-101 clips of 16 frames, one window per source video, at the 128×128 resolution both embeddings read (Ge et al., 2024). These 2048 clips keep the covariance nondegenerate in the 1408 dimensions of VideoMAE-v2. The distortions are the elastic transformation and motion blur of ImageNet-C at its five severities. We apply them at 128 pixels with the suite's pixel constants, which are set for larger images, so they are harsher relative to the frame than in ImageNet-C. FVD uses the 400-dimensional logits of the Kinetics-400 I3D classifier. The VideoMAE-v2 features are the 1408-dimensional outputs of its ViT-g model fine-tuned on Something-Something-v2, following Ge et al. (2024). The bands for the flicker set at one severity average the displacements of the other four severities before taking eigenvectors. The bands for the generator average all five severities. ModelScope generates 2048 clips of 16 frames from the UCF-101 class names with its default DPM-Solver scheduler at 25 steps, cropped like the reference.

Proteins. La-Proteina generates unconditional backbones from its released checkpoint. Each setting has 467 structures, the size of the Protein FID test set, which its default sampler spreads evenly over lengths of 100, 200, 300, 400, and 500 residues. The embeddings are ESM3 structure-token embeddings, mean-pooled to 1536 dimensions with the sequence masked. The reference is the released embeddings of 4943 curated PDB structures of Faltings et al. (2025), each a FoldSeek cluster representative. The PCA to 32 dimensions is fit jointly on the reference and generated embeddings, as in Protein FID, and we recompute the optimal transport in this space. The concept dictionary is built from the $31{,}885$ domains of the CATH S40 non-redundant set (Sillitoe et al., 2021). These domains are embedded the same way and projected onto the same PCA basis. The alpha-versus-beta direction runs from the mainly-alpha class to the mainly-beta class, and each geometric property is split at its terciles. The ten geometric properties are chain length, radius of gyration, globularity (the radius of gyration over the cube root of the length), the helix, strand, and coil fractions, the number of secondary-structure segments per residue, relative contact order, clashes per residue, and backbone-bond deviation, the standard deviation of consecutive alpha-carbon spacings.

D Diagnostics of the image directions

This appendix gives further diagnostics of the image directions for SD1.5 against COCO (Section 4.1). It separates the mean and covariance parts of their directional FD (Appendix B) and measures the composition-simplicity concept direction on its own.

The top eigenvector is the mean difference. e_1 has cosine 0.99 with the mean-difference direction at 50 steps. Its mean part is 85% of $\mathrm{FD}(e_1)$. The largest single share of FID is thus the mean gap between the two sets, laid along one direction. The next eigenvectors are almost all covariance part. Their mean part is 4% of $\mathrm{FD}(e_2)$ and under 1% of $\mathrm{FD}(e_3)$. Table 4 lists every concept of the sparse concept fits of Table 1. It also adds the fits that each dictionary gives to the eigenvectors the

Table 4: Every concept whose weight rounds to at least 0.01 in the sparse concept fits of the top eigenvectors of M, for SD1.5, 50 steps, and COCO. Each weight is positive when the concept, or class logit, grows toward $+e_i$, the generated end. The last column is the reconstruction cosine.

	dictionary	surviving concepts (weight)	cos
e_1	CLIP	composition simplicity ($+0.55$), depth of field ($+0.12$), furniture ($+0.08$), electronics ($+0.05$), daytime ($+0.04$), portrait versus scene (-0.03)	0.74
e_1	ImageNet	notebook ($+0.02$)	0.16
e_2	CLIP	brightness (-0.15), weather (-0.12), crowd density ($+0.08$), indoor/outdoor ($+0.08$), depth of field ($+0.05$), electronics ($+0.01$), centered composition (-0.01), vehicles ($+0.01$)	0.49
e_2	ImageNet	hand blower ($+0.08$), horse cart (-0.06), china cabinet ($+0.06$), Crock Pot ($+0.03$), space heater ($+0.02$), crane (-0.01), jinrikisha (-0.01), tank (-0.01)	0.43
e_3	CLIP	indoor/outdoor ($+0.03$), sports equipment ($+0.01$)	0.17
e_3	ImageNet	Dungeness crab (-0.09), mashed potato (-0.06), burrito (-0.05), spiny lobster (-0.02), orange ($+0.02$), ice cream (-0.02)	0.40

other dictionary names. The ImageNet-class dictionary reconstructs e_1 at cosine 0.16, and the CLIP dictionary reconstructs e_3 at 0.17.

The concept direction on its own. The name of e_1 rests on a sparse concept fit. We therefore also measure the composition-simplicity concept direction itself, which is built without reference to M. It has cosine 0.67 with e_1 and carries 10.9% of FID at 50 steps. Its directional FD has slope 0.39 against the total FID across the step sweep of Section 4.1.2, and its mean part accounts for 0.36 of this slope. The concept direction alone thus carries about two fifths of the FID increase. e_1, whose fit it dominates, carries two thirds.

E SIGN REVERSAL ACROSS REFERENCES

Section 4.1.2 decomposes the step-driven FID change on COCO into a rise over the head and a fall over the tail, where the rise outweighs the fall and FID rises with steps. The sign of this change is known to differ by reference (Lee & Pak, 2026), so we repeat the COCO sweep on six references and ask what sets the sign. The references are COCO30K, Flickr30K (Young et al., 2014), ImageNet (Deng et al., 2009), MJHQ30K (Li et al., 2024), CelebA-HQ (Karras et al., 2017), and FFHQ512 (Karras et al., 2019), spanning caption and object sets (COCO, Flickr, ImageNet), an aesthetic set (MJHQ), and two curated face sets. Each reference is compared with SD1.5 generations from its own captions, so the six comparisons differ in the generated set as well as in the reference. We compute each reference's own eigenbasis of M at 50 steps and hold it fixed throughout its sweep, as on COCO. The cumulative slope $s_{1:i}$ of Section 4.1.2 ends at 1 on every reference, so it cannot compare the total change across references. We instead regress each $\mathrm{FD}(e_i)$ against the step count, and the cumulative sum $\beta_{1:i} = \sum_{j \le i} \beta_j$ of the resulting slopes β_i ends at the total slope β_{FD} of FID against steps (Figure 4).

Rise and fall across references. The cumulative curves separate two effects that every reference shares in kind but not in size. The first is the head, led by e_1, along which the generated mean drifts as steps increase. The second is the tail, where directional FD falls with steps on every reference. The drift along e_1 runs away from the reference on COCO, Flickr, ImageNet, and MJHQ, so their curves rise over the head before falling over the tail. The drift runs toward the reference on the two face references, so their curves fall throughout, with the decrease spread across the spectrum. The tail fall is of comparable size on every reference, so the head decides whether FID reverses. FID rises when the head rise outweighs the tail fall, falls when the drift runs toward the reference and the head falls too, and stays nearly flat when the two cancel.

Aligned drifts on reference-specific eigenvectors. The change in the generated mean across the sweep, $\mu_g(50) - \mu_g(15)$, aligns across all six comparisons (pairwise cosine 0.52 to 0.88), although their captions differ. Raising the step count thus moves the generated mean in a similar direction whatever the reference. Each reference's FID, however, is carried by its own head, so the same drift

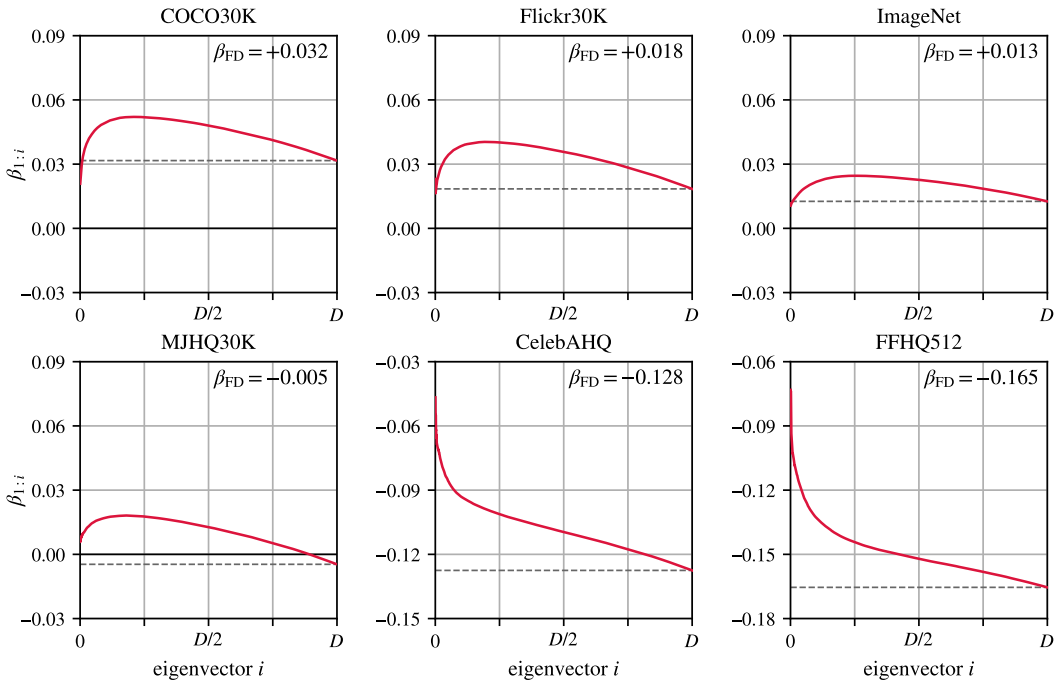


Figure 4: Cumulative slope $\beta_{1:i} = \sum_{j \le i} \beta_j$ of the top i fixed eigenvectors against i. Here β_i is the slope of the directional FD $\mathrm{FD}(e_i)$ against the step count, in each reference's own fixed 50-step basis. The curves of COCO, Flickr, ImageNet, and MJHQ rise over the head before falling over the tail, while the face curves descend throughout. Each endpoint is the total slope β_{FD}, which is positive for COCO, Flickr, and ImageNet, negative for the faces, and near zero for MJHQ.

contributes differently to each FID. The generated mean moves away from the reference along e_1 on COCO, Flickr, and ImageNet, where the head rise exceeds the tail fall and FID worsens. The mean moves toward the reference along e_1 on CelebA-HQ and FFHQ, and directional FD also decreases broadly across their spectra, so FID improves. MJHQ is the cancellation case: its discrepancy along e_1 grows like that of the caption and object references, but the tail fall offsets it and leaves the total FID nearly unchanged.

Finding. Adding steps drives the generated mean in a similar direction in every comparison. Each reference's FID, however, is carried by its own head, so the reference decides whether FID reverses. The caption and object references (COCO, Flickr, ImageNet) worsen, the face references improve, and MJHQ cancels.

F VIDEO BANDS

Overlap between the bands. The appearance and flicker bands are computed independently from $M(R, S)$ and $M(S, T)$, so they need not be orthogonal. The two shares of a row of Table 2 therefore measure how much of a distance aligns with each band, not an exclusive allocation between them. With elastic, the two shares of I3D at $k = 10$ even sum past one. Table 5 measures this overlap at $k = 10$. It reports the largest principal-angle cosine between the two bands and the share of their union. It also reports the share of each band after orthogonalizing it against the other, which equals the union's share minus the other band's share. The ordering between the two bands on the flicker set holds for both embeddings and in both orders of orthogonalization with motion blur. The elastic bands keep the ordering too, except for I3D when the appearance band is orthogonalized against the flicker band.

Table 5: Overlap of the $k = 10$ appearance and flicker bands. The flicker set is $M(R, T)$ at severity 3 with bands computed from the other severities, and the generated set is $M(R, G)$ with bands computed from all severities. Columns give the largest principal-angle cosine between the bands, the share of their union, and the share of the flicker band orthogonalized against the appearance band (flicker | appearance) and the reverse.

embedding, family	set	cos	union	flicker | appearance	appearance | flicker
I3D, motion blur	flicker set	0.74	0.746	0.083	0.473
I3D, elastic	flicker set	0.80	0.833	0.185	0.338
VideoMAE, motion blur	flicker set	0.61	0.786	0.605	0.174
VideoMAE, elastic	flicker set	0.69	0.741	0.459	0.188
I3D, motion blur	generated	0.74	0.384	0.096	0.234
I3D, elastic	generated	0.80	0.371	0.083	0.265
VideoMAE, motion blur	generated	0.59	0.183	0.088	0.087
VideoMAE, elastic	generated	0.68	0.205	0.056	0.121

Table 6: Scalar ratio and top flicker-eigenvector share across the five motion blur severities, with bands computed from the other severities.

severity	1	2	3	4	5
I3D ratio	1.41	1.29	1.18	1.11	1.07
VideoMAE ratio	3.56	3.63	3.23	2.87	2.77
I3D flicker share, $k = 1$	0.053	0.028	0.023	0.028	0.031
VideoMAE flicker share, $k = 1$	0.235	0.371	0.470	0.475	0.374

The generator. The shares on the generator are smaller and closer than on the flicker set. The two bands of VideoMAE are tied with motion blur at $k = 10$, in both orders of orthogonalization. What VideoMAE changes relative to I3D is thus the balance between the bands. The ratio of the flicker share to the appearance share in Table 2 rises from 0.52 under I3D to 1.01 under VideoMAE. The elastic ratio rises from 0.37 to 0.57, and appearance still leads there. The union of the two bands holds only 18% to 38% of the generated distance at $k = 10$. Most of the generator's discrepancy from UCF-101 lies outside the directions that the distortions define.

Across severities. Table 6 compares two quantities at every severity of motion blur. The first is the scalar ratio of Ge et al. (2024), the flicker set's FVD over the appearance set's. The second is the share of the flicker set's FVD on the top flicker eigenvector. The ratio falls toward one as severity grows for both embeddings. The share keeps the two embeddings apart at every severity.

G PROTEIN FID IN THE RAW EMBEDDING SPACE

The main text analyzes the 32-dimensional PCA space that defines Protein FID. This appendix repeats the analysis in the raw 1536-dimensional ESM3 space and lists the full sparse concept fits in both spaces.

The same directions. The raw FD is 3.40×10^6, against 2.87×10^6 for Protein FID. The raw directional FD of the PCA-32 subspace is 2.90×10^6, so the lower bound of Appendix B is nearly tight. The raw spectrum of M is less steep: e_1 carries 56% of the raw FD, e_2 12%, and e_3 4% (Table 7). The PCA thus keeps the dominant discrepancy and makes it slightly more concentrated. The raw e_1 reconstructs at cosine 0.71 and is dominated by a coil deficit. The generated set is shifted along it by Cohen's $d = 3.2$, as in the PCA space. The reference varies along e_1 more than along all but two of its principal directions in both spaces. The lower eigenvectors keep their fits and their narrow generated spread. The alpha-versus-beta direction carries 4.5% of the raw FD with cosine 0.23 to the mean shift, against 5.5% and 0.24 for Protein FID.

Table 7: Full sparse concept fits of the top eigenvectors of M for La-Proteina at 400 steps, in the 32-dimensional PCA space of Protein FID (Table 3, penalty 10^{-3}) and in the raw 1536-dimensional ESM3 space (penalty 10^{-5}). Every surviving concept is listed with its weight, and the other columns are as in Table 3.

space		share	surviving concepts (weight)	cos	d	var
PCA-32	e_1	66%	coil (-0.97), contact order ($+0.53$), length ($+0.16$), globularity (-0.16), radius of gyration (-0.10)	0.72	$+3.2$	0.55
	e_2	14%	backbone-bond deviation (-0.58), clash ($+0.50$), contact order ($+0.18$), segments per residue ($+0.17$)	0.92	$+0.0$	0.05
	e_3	4%	contact order (-0.90), clash ($+0.31$), segments per residue ($+0.18$), backbone-bond deviation ($+0.17$), radius of gyration ($+0.02$)	0.83	$+0.1$	0.16
raw	e_1	56%	coil (-0.99), length ($+0.51$), contact order ($+0.45$), radius of gyration (-0.36), alpha-versus-beta ($+0.06$), segments per residue (-0.04)	0.71	$+3.2$	0.56
	e_2	12%	backbone-bond deviation (-0.58), clash ($+0.50$), contact order ($+0.19$), segments per residue ($+0.18$), strand ($+0.06$), coil (-0.06), globularity ($+0.04$)	0.90	$+0.0$	0.04
	e_3	4%	contact order (-0.82), clash ($+0.30$), backbone-bond deviation ($+0.17$), segments per residue ($+0.13$), radius of gyration ($+0.02$), strand (-0.01), length (-0.01)	0.76	$+0.1$	0.16

Correlated concepts. The concept directions are strongly collinear in the PCA space. The alpha-versus-beta direction has cosine -0.98 with the helix direction, and helix has -0.96 with coil. At the raw-space penalty of 10^{-5}, the PCA-space fit therefore trades large opposite weights among the helix, strand, and coil fractions, which sum to one. We raise the penalty to 10^{-3} in the PCA space, where every weight stays below one. Even so, a weight can differ from the alignment of its concept direction alone. Contact order enters both fits of e_1 with a positive weight, although its direction alone has cosine -0.10 with e_1 in either space. This weight is a correction among correlated concepts, not a higher contact order at the generated end.