RETHINKING ZERO-SHOT CAPABILITIES IN VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision foundation models are used zero-shot on tasks they were never trained for, such as matching features across images for semantic and geometric correspondence, tracking points through video, and nearest-neighbour classification. Evaluations extract features, compare them by cosine similarity, and choose the candidate with the highest score. Cosine similarity does not account for how variance is distributed across feature directions. In many models, a few directions carry most of the variance. We measure this concentration by the effective dimension $d_{\text{eff}} = (\operatorname{Tr}\boldsymbol{\Sigma})^2/\operatorname{Tr}(\boldsymbol{\Sigma}^2)$ of the covariance $\boldsymbol{\Sigma}$ of the extracted feature vectors. Across 36 models and 46 extraction points, d_{eff} ranges from 1.7 to 374, despite feature dimensions ranging from 768 to 6144. Such low effective dimensions matter because the cosine similarity between unrelated features has variance of about $1/d_{\text{eff}}$, so when d_{eff} is small, wrong candidates can score almost as highly as the correct match. We therefore whiten the features with their mean and covariance, estimated on unlabeled images, before computing cosine similarity. Whitening requires no labels or training and has long been standard in image retrieval, yet it is not part of the usual zero-shot evaluation of foundation-model features. Across eleven dense benchmarks, whitening improves performance on 459 of 506 model–benchmark pairs, by 5.4 points on average and up to 45 PCK points. The gain is larger for models with lower pre-whitening d_{eff}. Diffusion transformers gain the most. Their raw features match poorly under cosine similarity, but after whitening they achieve matching performance comparable to strong self-supervised encoders. Whitening should be part of the zero-shot evaluation of frozen vision features.

1 INTRODUCTION

Frozen vision foundation models (FMs) represent an image either with dense feature vectors, one for each image patch (a small square of pixels), or with a global image feature obtained from a special token or by pooling patch features. Dense features are typically used for pixel-level tasks, while global features are used for image-level tasks. These features can be used zero-shot to match patches between images, transfer keypoints, estimate pose, track points through video, and classify images by nearest neighbour, without task-specific training. The usual procedure is to extract the features, compare them by cosine similarity, and choose the candidate with the highest score (Tang et al., 2023; Zhang et al., 2023; Oquab et al., 2024; Siméoni et al., 2026). Cosine similarity does not account for how variance is distributed across feature directions. In many FMs, a few directions carry most of the variance (Section 3). The degree of this concentration varies across FMs by more than two orders of magnitude, raising the question of how much the choice of similarity metric affects the observed performance. We study this effect and show that whitening the features with their own mean and covariance can substantially change the evaluation.

Dense correspondence makes the problem concrete. Consider a query patch in one image and the K patches of a second image, where K is typically 10^3 to 10^4 and exactly one patch is the true match. Cosine similarity between the query and an unrelated candidate is not zero; it fluctuates from candidate to candidate, and the true match must score above the largest of these $K-1$ wrong-match scores. The spread of these scores depends on the model's feature distribution. We summarise this concentration with an effective dimension d_{eff}, which is large when variance is spread across many feature directions and small when it is concentrated in a few. For unrelated features, the

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



Figure 1: **One query, six readouts.** The white point is matched into a different bus; panels are ordered by each readout's d_{eff}. The four raw diffusion transformers miss under cosine similarity (red) and match the point after whitening (teal). The two encoders have high d_{eff} and are shown for comparison, and they are correct either way. The DiTs are those with SPair-71k accuracy below 50% under cosine and near or above it after correction (Table 1). This example is selected. Of the 426 keypoints in a random draw of 60 test pairs, 20 show this pattern, for 90 whitening changes no readout's answer, and for 83 at least one readout gets worse. Two more examples are in Appendix D.

variance of cosine similarity is approximately $1/d_{\text{eff}}$ (see Section 3 and Appendix B.1). Thus, when d_{eff} is small, the scores of wrong candidates spread more widely, making the highest score among thousands of candidates large enough to compete with the true match (Figure 1). We measure d_{eff} for 46 readouts of 36 models, where a readout is a feature extracted at one point in the network; diffusion transformers are read at two points, before and after a block's normalisation (Appendix A). d_{eff} ranges from 1.7 to 374 for feature dimensions ranging from 768 to 6144. One direction carries two thirds of CLIP-B's patch variance (Radford et al., 2021), while DINOv2-L (Oquab et al., 2024) requires hundreds of directions to account for the same fraction (Figure 2). Raw cosine similarity therefore evaluates models with very different feature geometries.

We address this problem by estimating the mean and covariance of the patch features on unlabeled images and whitening the features with them before computing cosine similarity, as in image retrieval (Jégou & Chum, 2012) and dense HOG matching (Hariharan et al., 2012; Bristow et al., 2015). Across eleven benchmarks, whitening improves performance on 459 of 506 model–benchmark pairs. The gain is largest for readouts with the smallest d_{eff} (Figure 5), and whitening changes the ranking of models (Table 1). We therefore compare frozen features after whitening.

Diffusion transformers are the clearest example. Their features can be read at two points of the same block, before or after the block's normalisation layer. We call the reading from the residual stream before it the raw readout. The two differ by a per-channel affine map, so they carry nearly the same information. Under cosine similarity the raw features match far worse. On SOCO the post-normalisation readout scores up to 41 points higher (Dünkel et al., 2026), a gap that Gan et al. (2025) attribute to a few massive-activation channels and remove architecturally. Averaged over the ten diffusion transformers in our study, cosine separates the two readouts by 15.4 points; after whitening they agree to within 1.1 (Figure 5). On SPair-71k (Min et al., 2019), raw FLUX.1 (Labs, 2024) rises from 16.8 to 61.8, from 38th to 3rd of 46 readouts.

Contributions.

1. **A diagnosis.** We derive that the cosine similarity between unrelated frozen features has variance approximately $1/d_{\text{eff}}$, where d_{eff} is the effective dimension of the feature distribution. Thus, when d_{eff} is small, some wrong candidates can score almost as highly as the true match. We verify this prediction on real features, and find that d_{eff} varies across models by two orders of magnitude (Section 3).
2. **A fix.** We whiten the features using their mean and covariance, estimated from unlabeled images, before computing cosine similarity. This requires no training or labels and improves performance on 459 of 506 model–benchmark pairs (Sections 4 and 5).

3. **A consequence.** Whitening changes the relative performance and ranking of models, showing that cosine similarity can substantially underestimate some representations. Across ten diffusion transformers, whitening makes even raw (pre-normalisation) features competitive with strong self-supervised encoders, with some surpassing them. We therefore recommend whitening as a standard step in the zero-shot evaluation of vision foundation models.

2 Related work

Whitening before nearest-neighbour search. Jégou & Chum (2012) established PCA-whitening as a standard post-processing step in image retrieval. In NLP, removing the dominant directions of an embedding improves similarity (Mu et al., 2018; Su et al., 2021), Timkey & Van Schijndel (2021) repair cosine by standardising the coordinates responsible, and Jung et al. (2023) whiten the token embeddings of a multi-vector retriever. Before deep features, Hariharan et al. (2012) and Bristow et al. (2015) whitened dense HOG features using statistics from unlabeled natural images. Other work learns the transform from matching supervision (Radenović et al., 2018), introduces decorrelation during training (Ermolov et al., 2021; Zbontar et al., 2021; Bardes et al., 2022), adds a whitening layer to the encoder (Kalapos & Gyires-Tóth, 2024), or adjusts similarity scores rather than the features (hubness, Radovanović et al., 2010; Conneau et al., 2018; Bogolin et al., 2022). We use the unsupervised form, with no training or labels, on global and dense features from frozen vision foundation models, and study its effect on zero-shot evaluation.

Anisotropy. Anisotropic representations concentrate their variance in a few directions, so the features occupy a narrow cone and unrelated items can already have high similarity. Learned embeddings are strongly anisotropic (Ethayarajh, 2019; Gao et al., 2021). This has been linked to a small number of outlier coordinates (Kovaleva et al., 2021; An et al., 2025) and measured in several ways (Cai et al., 2021; Rudman et al., 2022), including d_{eff}. Steck et al. (2024) show that cosine similarities are affected by diagonal rescaling, and Parupudi (2026) find across 19 encoders that anisotropy affects whether cosine or a rank metric works better. In vision, anisotropy has been studied at the image level (Godey et al., 2023) and through unusual token norms that motivate registers (Darcet et al., 2024). Kang et al. (2025) prove that no CLIP-like joint space satisfies several vision–language desiderata under cosine, but they consider one image embedding against text and use a trained module as the remedy. A similar issue appears in dense vision features. Siméoni et al. (2026) document channels whose magnitudes grow with depth and remain high *across patches and images*, unlike the input-dependent outliers reported for language models (An et al., 2025). A channel that is large for every patch cannot distinguish one patch from another, yet it contributes strongly to every cosine score. We examine this effect across 46 readouts from 36 modern vision foundation models.

Correspondence pipelines. Several correspondence pipelines use additional feature processing before computing similarity. Zhang et al. (2023) L_2-normalise each branch and reduce it by PCA before concatenating Stable Diffusion with DINO features, a step Zhang et al. (2024) inherit. Gan et al. (2025) attribute poor matching of raw diffusion-transformer features to a few channels with very large activations (as in language models, Sun et al., 2024) and suppress them using the model's AdaLN modulation together with explicit channel removal. These operations modify the feature geometry in different ways. PCA projects without rescaling, channel removal sets selected coordinates to zero, and AdaLN applies a diagonal scale supplied by the model rather than estimated from the feature distribution. Our whitening transform is estimated directly from feature statistics, so it does not require access to model internals or a particular normalization or modulation mechanism.

Our contribution is not whitening itself or the observation that feature distributions can be anisotropic. We systematically measure how much this effect varies across modern vision foundation models. Across 46 readouts of 36 models, d_{eff} varies by two orders of magnitude. These differences matter because model comparisons based on frozen features use cosine similarity as part of the evaluation. Zhang et al. (2025) conclude that backbone choice dominates matching machinery, and Dünkel et al. (2026) that correspondence predicts dense downstream ability better than ImageNet classification; both compare backbones using cosine. Earlier work on whitening focuses on improving similarity. We use whitening to show how much the choice of similarity measure can change the evaluation of frozen vision features.

3 WHAT COSINE MEASURES

A frozen backbone maps an image to a grid of feature vectors $\mathbf{x}_p \in \mathbb{R}^d$, one per patch (Figure 3). To find the match of a query patch p in a target image, zero-shot evaluation compares $\mathbf{x}_p$ with the feature $\mathbf{y}_q$ of every patch q in the set $\mathcal{Q}$ of the K target patches and takes the most similar one,

$$\hat{q} \;=\; \arg\max_{q \in \mathcal{Q}} \cos(\mathbf{x}_p, \mathbf{y}_q), \qquad K = |\mathcal{Q}| \sim 10^3\text{–}10^4. \tag{1}$$

The prediction counts as correct if it lands within a benchmark-specific distance of the true location (Section 5). Nothing is trained, so the score is read as a property of the features. We show that the score also depends on how the feature dimensions are scaled, and measure how much.

The true match competes with $K-1$ wrong candidates, so we first ask what a wrong candidate scores. For two patches drawn independently from the feature distribution, the cosine averages $\|\boldsymbol{\nu}\|^2$, where $\boldsymbol{\nu} = \mathbb{E}[\mathbf{x}/\|\mathbf{x}\|]$ is the mean direction of the features. After subtracting this mean score, the variance of the cosine is (Appendix B.1)

$$\mathrm{Var}\big(\cos(\mathbf{x},\mathbf{y})\big) \;\approx\; \frac{\mathrm{Tr}(\boldsymbol{\Sigma}^2)}{(\mathrm{Tr}\,\boldsymbol{\Sigma})^2} \;=\; \frac{1}{d_{\text{eff}}}, \qquad d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \tag{2}$$

where λ_i are the eigenvalues of $\boldsymbol{\Sigma}$, the covariance of the patch features. d_{eff} is the participation ratio (Gao et al., 2017). It equals d when all directions carry the same variance and 1 when a single direction carries all of it. Cosine measures an angle, and a direction in which the features barely vary hardly changes any angle, whatever it encodes. When a few directions hold most of the variance, all patches point roughly the same way, the cosines of unrelated pairs cluster around one value, and the true match has little room to stand out (Figure 4, left).

Figure 2 shows how real models distribute their variance. A curve that rises steeply on the left has its variance in a few directions and a small d_{eff}. Raw diffusion-transformer readouts reach two thirds of their variance within a handful of directions (6 for FLUX.1), vision encoders within a few hundred (258 for DINOv2-L), and the same diffusion weights after their own normalisation only after several hundred (411 for FLUX.1). d_{eff} summarises the same curve differently, weighting each direction by its variance, so it does not equal these counts. Across the 46 readouts it ranges from 1.7 to 374 in 768 to 6144 dimensions (Figure 6, Appendix B.2). It depends more on where the features are read than on the model (FLUX.1: 3 raw, 289 after AdaLN).

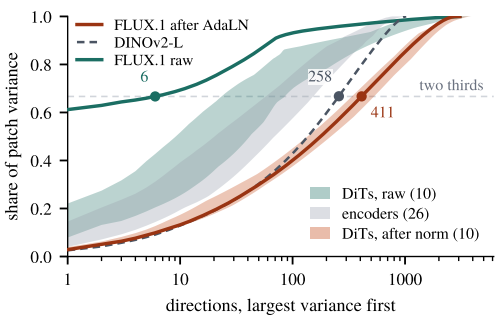

Figure 2: **Where the variance sits.** Cumulative share of a readout's patch variance in its k largest principal directions. Bands: interquartile range per family. Lines: FLUX.1 raw and after AdaLN (same weights), and DINOv2-L. FLUX.1 raw and DINOv2-L sit near the ends of their families, outside the bands. Dots: two thirds of the variance.

For matching, what counts is not a typical wrong candidate but the best of $K-1$ of them. For $K-1$ independent scores with variance $1/d_{\text{eff}}$, the expected maximum is at most $\sqrt{2\ln K/d_{\text{eff}}}$ (Appendix B.3, Figure 11). For FLUX.1 after AdaLN and $K = 10^4$ this bound is 0.25, far below 1. For the same weights read raw it is 2.6. The bound is then empty, wrong candidates crowd against 1, and which patch wins is close to chance.

Diffusion transformers let us test this directly. Inside each block, the residual stream is normalised and every channel is multiplied by a scale computed from the timestep and prompt. We read the features either before this step (raw) or after it (post-normalisation). The two differ only by a per-channel rescaling and contain nearly the same information, yet under cosine they score up to 41 points apart on SOCO (Figure 5). Gan et al. (2025) attribute the poor raw scores to a few channels with very large activations and remove them. Following them, our post-normalisation readouts also remove these channels (Table 4). Whitening still improves these readouts on 76 of 110 evaluations, so the cosine scores depend on more than those channels. Rescaling the channels alone can move d_{eff} to any value between 1 and d (Appendix B.5). Whitening (Section 4) makes the score largely independent of how the channels are scaled.

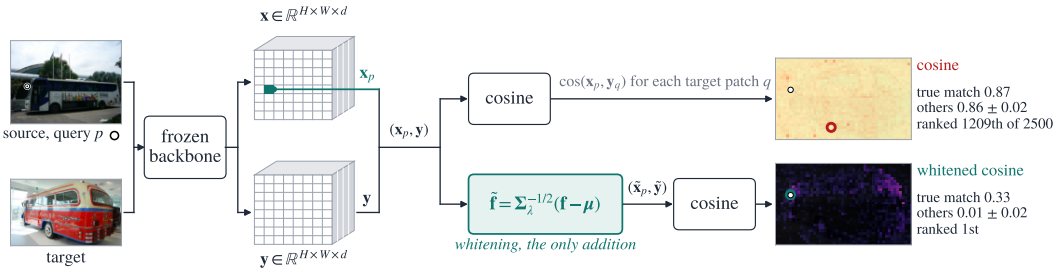

Figure 3: **Same features, two metrics.** Both images pass through the same frozen backbone. The query feature (white ring, teal cell) is compared with every target patch by cosine similarity (Eq. (1)), directly or, in the lower path, after whitening (Eq. (3)). The maps show the scores for the query of Figure 1 on FLUX.1 read raw, each divided by its maximum and overlaid on the target. Only the ranking matters, so what counts is how far the true match stands above the other patches. Under cosine it scores the same as the other 2499 patches and ranks 1209th. After whitening it is the only peak. Whitening leaves the true match high and pushes the other patches down.

4 Correcting the metric

Section 3 traced the problem to two properties of the features. Their variance is concentrated in a few directions, and they sit away from the origin, so even unrelated patches have a high expected cosine, $\|\boldsymbol{\nu}\|^2$. The correction handles both. Subtracting the mean $\boldsymbol{\mu}$ moves the features to the origin and brings this expected cosine close to zero (Figure 4b). Multiplying by $\boldsymbol{\Sigma}^{-1/2}$ turns the covariance into the identity, so every direction carries the same variance and d_{eff} reaches its maximum d, which by Eq. (2) minimises the spread of wrong-candidate scores (Figure 4c). For a patch feature $\mathbf{f} \in \mathbb{R}^d$, with $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}}$ the mean and covariance of patch features over a set of unlabeled images, we use

$$\tilde{\mathbf{f}} \;=\; \boldsymbol{\Sigma}_\alpha^{-1/2}(\mathbf{f} - \boldsymbol{\mu}), \qquad \boldsymbol{\Sigma}_\alpha = (1-\alpha)\hat{\boldsymbol{\Sigma}} + \alpha\bar{\sigma}^2 I, \quad \bar{\sigma}^2 = \tfrac{1}{d}\operatorname{Tr}\hat{\boldsymbol{\Sigma}} \tag{3}$$

and then compute cosine as before, with $\cos(\tilde{\mathbf{x}}_p, \tilde{\mathbf{y}}_q)$ in place of $\cos(\mathbf{x}_p, \mathbf{y}_q)$ in Eq. (1) (Figure 3).

The model itself is not changed. There are no labels or gradient steps, and the backbone, extraction point and matching rule stay the same. $\boldsymbol{\Sigma}_\alpha^{-1/2}$ is computed once per readout from one $d \times d$ eigendecomposition. At match time the correction costs one subtraction and one matrix multiplication per feature, small next to the backbone's forward pass. All experiments run on a single NVIDIA A6000 (48 GB).

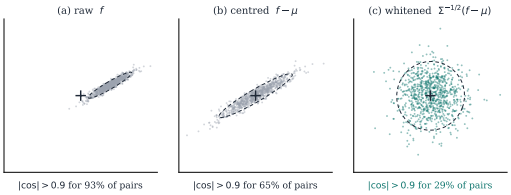

Figure 4: **What centring and whitening each fix** (synthetic 2-D features). The cross marks the origin, from which cosine measures angles. Below each panel is the fraction of random pairs with $|\cos| > 0.9$. Centring lowers the expected cosine $\|\boldsymbol{\nu}\|^2$ (93% → 65%), and whitening removes the remaining excess spread (65% → 29%).

The baseline is a special case of Eq. (3). With $\alpha = 1$, $\boldsymbol{\Sigma}_\alpha = \bar{\sigma}^2 I$ scales every feature by the same factor, which leaves all angles unchanged. If we also skip the subtraction of $\boldsymbol{\mu}$, the result is plain cosine. In every configuration this setting reproduces the baseline to floating-point precision, so the two scores differ only in the metric. Keeping the subtraction gives centred cosine, and $\alpha = 0$ gives full whitening. Replacing $\hat{\boldsymbol{\Sigma}}$ by its diagonal rescales each channel and ignores correlations. Section 5.3 shows that this already gives most of the gain.

The shrinkage parameter α (Ledoit & Wolf, 2004; Friedman, 1989) controls how far the correction goes towards full whitening. Full whitening gives equal weight to low-variance directions that may carry little signal, and divides by the smallest eigenvalues of $\hat{\boldsymbol{\Sigma}}$, which are the least reliable when few images are used. We select α once per readout on a validation split and use the same value for every benchmark (Section 5). Other rules for α, from a fixed value to a per-benchmark oracle, rank encoders, raw and post-normalised diffusion transformers in the same order (Table 6, Appendix C).

5 EXPERIMENTS

Models and readouts. We evaluate 36 frozen models at 46 readouts. The 26 vision encoders are read once. The 10 diffusion transformers are read twice, at the input to a block (the residual stream, *raw*) and after that block's AdaLN or RMSNorm. For the post-normalisation readout we also remove the massive-activation channels listed in Table 4, as DiTF does (Gan et al., 2025), so this readout already includes their fix. We use published extraction settings where they exist, namely the final patch tokens for most encoders, the intermediate layer given in the model's paper where there is one, the setting of Tang et al. (2023) for the Stable Diffusion U-Nets, and that of Gan et al. (2025) for FLUX.1, SD3.5-Large and PixArt-α. The other seven diffusion transformers have not been evaluated as dense feature extractors before, so there is no established block or timestep. For these we choose both on the SPair-71k validation split by maximising the *cosine* score, and keep them fixed for every benchmark (Table 4). Each diffusion model is run with a single noise sample, as SOCO does for SD2.1 and as an encoder gets a single forward pass. DIFT and DiTF average eight samples, which would raise the baselines. α is chosen the same way, once per readout on the SPair-71k validation split, and used unchanged on every benchmark, including the geometric and tracking ones. No test split is used for any choice. Because extraction settings maximise the cosine score, our gains are measured against the best cosine baseline for each model.

Estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Unless stated otherwise, the mean and covariance in Eq. (3) are computed from the patch features of 250 images from the ImageNet training set (Deng et al., 2009), which share no images or labels with any benchmark. This gives 1.6 to 8.1×10^5 patch vectors per readout, 38 to 1058 times the feature width d. Redrawing the 250 images changes the gain by at most 0.4 points (Appendix E). Section 5.3 shows that a single image is enough, and that the moments can even be taken from the image being matched.

Benchmarks and metrics. We use OmniProbe, the evaluation suite of SOCO (Dünkel et al., 2026), and follow its zero-shot protocols. It has eleven dense evaluations over seven datasets, none with a trained head. Four come from SOCO itself. Semantic object correspondence (SOC) and concept correspondence (CC) use its 20k within-category pairs, cross-category correspondence (Cross-SOC) uses its 20k cross-category pairs, and in SOC-Geo the candidates are restricted to keypoints of the same concept, so that only geometry (left leg or right leg) decides. The others are SPair-71k (Min et al., 2019), the standard semantic benchmark; AP-10K (Yu et al., 2021), animal keypoints over fifty species, on the intra-species, cross-species and cross-family test splits of Zhang et al. (2024) with 5000 pairs each; ScanNet (Dai et al., 2017), indoor RGB-D video, on the SuperGlue test split of 1500 pairs with 1000 sampled correspondences per pair; NAVI (Jampani et al., 2023), multi-view images of 36 objects with aligned meshes; and TAP-Vid DAVIS (Doersch et al., 2022), point tracking through video. The metric is per-image PCK@0.1 of the bounding box on the semantic benchmarks, recall within 10 px on ScanNet, recall within 2 cm in 3D on NAVI and average Jaccard on TAP-Vid (Appendix A). We report differences in *points* of each benchmark's own measure. Table 1 shows six of the eleven evaluations and Appendix D the other five. We also run k-NN classification on ImageNet as in OmniProbe, with top-1 accuracy, the CLS token where a model has one and the mean of the patch tokens otherwise, and k fixed at 20 (Section 5.1). Its $d_{\text{eff}}^{\text{cls}}$ in Table 1 is computed on that global vector.

5.1 MAIN RESULTS

Table 1 shows every readout on six of the eleven evaluations, before and after whitening. Of the 506 model–benchmark pairs, 459 improve, 45 get worse and 2 are unchanged. Raw DiT readouts improve on all 110 of their evaluations, by $+14.0$ on average. Encoders improve on 273 of 286, by $+3.6$, and post-normalised DiT readouts, which have the highest d_{eff} in the study, by $+1.4$. The largest gains are on the readouts that cosine scored worst. The gain also holds across categories. On AP-10K it is $+4.9$ within species, $+5.6$ across species and $+4.8$ across families (Appendix D). Choosing α on validation costs little, giving 93% of the gain of a per-evaluation oracle and 98–99% on every semantic evaluation except SOC-Geo (Appendix C).

The gain is not specific to matching patches. On k-NN classification on ImageNet, which compares one global vector per image and involves no patches, whitening improves 37 of 46 readouts by $+5.2$ points on average, $+6.3$ on diffusion transformers and $+4.4$ on encoders (Table 11). Of the eight that

			semantic correspondence			geometric correspondence		tracking		classification
	readout	d_{eff}	SOC	SPair-71k	AP-10K	ScanNet	NAVI	TAP-Vid	$d_{\text{eff}}^{\text{cls}}$	kNN-IN
vision encoders	DINOv2-L/14	277	62.1→62.6	57.2→58.4	62.3→62.8	27.1→26.8	71.8→72.3	24.1→24.2	566	83.5→83.5
	DINOv3-L/16	143	61.0→62.0	56.5→60.0	61.3→62.1	51.1→55.7	77.1→78.8	21.9→22.7	460	85.3→85.3
	DINOv2-B/14	233	60.5→60.8	56.5→57.7	61.7→62.1	27.5→29.8	68.9→69.6	22.9→23.3	449	82.2→82.2
	DINOv3-B/16	128	58.6→59.5	55.7→58.8	60.3→61.3	48.9→53.3	73.6→76.5	22.1→22.6	358	83.0→83.0
	C-RADIOv3-B	157	51.1→51.5	49.4→50.0	52.0→52.2	34.4→37.1	66.2→67.2	19.4→19.7	93	77.8→77.6
	CleanDIFT	106	50.6→54.6	52.5→56.5	50.5→53.7	40.6→46.9	65.1→69.9	19.9→20.0	36	22.4→31.6
	SD2.1 t=250	149	49.0→52.2	50.9→54.2	49.1→51.4	42.5→47.5	57.7→63.0	15.4→16.0	46	5.8→12.3
	DUNE-B/14	33	46.7→49.9	43.0→48.0	48.1→51.2	27.2→44.5	64.0→68.8	24.8→25.6	49	57.8→60.4
	I-JEPA-H/16	69	46.3→50.1	41.2→47.1	48.0→51.6	11.7→18.6	55.4→58.5	17.6→17.9	104	61.1→62.9
	PE-Spatial-B	38	45.9→47.5	48.4→50.1	46.0→47.6	36.0→40.0	60.6→62.4	20.4→20.6	37	53.5→59.1
	DUNE-B/14$_{448}$	57	44.8→48.8	40.8→46.1	46.7→50.0	26.4→41.6	63.4→68.0	24.4→25.2	49	49.3→52.5
	SD2.1 (t=0)	107	44.8→48.7	46.4→52.5	47.2→51.2	47.6→58.2	61.3→65.3	18.9→19.8	27	12.0→23.3
	V-JEPA2.1-B	70	44.5→47.3	43.6→47.0	46.7→49.3	43.4→45.0	60.5→64.3	20.6→21.5	33	45.5→50.4
	iBOT-B/16	183	39.6→40.7	34.8→36.4	45.4→46.4	28.2→31.2	58.1→59.9	19.7→20.1	272	77.2→77.2
	DINO-B/16	107	30.6→32.5	26.7→29.0	33.2→35.5	34.0→37.5	56.6→58.6	19.5→20.1	207	76.1→76.1
	MiDaS-L/16	218	27.1→27.9	23.4→24.5	30.0→31.1	36.3→39.1	52.3→55.4	18.2→18.5	299	60.8→62.0
	SAM-B	39	25.5→28.3	24.9→28.9	26.1→29.6	43.0→46.2	54.9→56.6	14.3→18.4	12	26.4→40.4
	ConvNeXt-B	79	23.5→34.0	13.3→25.5	22.8→32.6	35.2→50.9	49.9→62.1	7.0→8.3	119	83.0→82.9
	DeiT3-B/16	39	23.4→26.8	16.2→21.3	22.0→25.7	14.9→20.2	37.0→40.4	12.8→14.1	15	84.2→84.5
	MetaCLIP2-B	2	20.2→22.6	19.9→22.2	21.9→23.9	13.1→18.4	36.9→41.4	13.3→13.9	92	77.3→76.6
	CLIP-B/16$_{\text{LAION}}$	2	16.7→19.8	16.2→20.2	19.8→22.6	14.1→17.4	36.0→38.1	12.7→13.5	100	76.9→77.2
	CLIP-B/16	2	16.2→21.1	15.9→21.5	21.1→25.7	17.1→25.6	35.2→39.9	12.6→14.0	100	73.9→73.7
	VGGT-1B	10	16.1→40.7	11.4→36.8	13.8→41.7	4.8→24.0	44.2→65.4	13.1→21.1	3	12.2→34.1
	CroCo-v2	11	10.2→14.1	8.4→11.9	10.4→13.6	23.4→37.5	45.4→47.7	14.3→16.1	8	20.2→33.4
	MAE-B/16	45	9.4→13.0	6.8→9.7	10.4→13.4	16.6→16.4	40.1→44.0	13.4→16.2	38	48.3→54.8
	SigLIP-B/16	5	8.8→11.2	6.7→8.9	9.8→12.6	12.2→15.6	25.9→30.3	7.9→8.9	8	14.5→26.5
	mean gain		+3.7	+4.5	+3.6	+6.5	+4.1	+1.2		+4.3
DiT, raw block	FLUX2-klein	65	58.6→62.8	54.5→62.9	59.2→62.7	44.0→47.9	71.3→74.5	20.1→22.4	94	66.9→69.0
	Z-Image	2	55.0→58.5	58.0→61.5	56.1→57.3	32.6→44.2	61.4→69.1	19.0→20.4	1	36.8→42.0
	ERNIE-Image	21	38.5→55.7	38.2→58.6	42.9→55.4	32.8→46.0	51.3→66.6	15.1→21.1	13	37.6→54.1
	SD3.5-L	49	34.3→52.0	35.3→56.5	40.5→53.4	30.3→48.5	52.4→67.5	14.1→20.9	7	15.4→32.6
	Qwen-Image	94	32.5→45.1	33.6→46.5	38.0→43.9	24.4→39.6	43.7→58.8	7.3→13.7	17	22.4→29.8
	Krea2	53	28.9→47.8	30.8→49.2	35.4→42.7	14.5→34.2	32.4→60.0	5.7→13.6	13	30.2→40.0
	PixArt-α	55	28.9→44.9	28.4→48.8	36.4→49.0	23.2→38.0	47.0→57.2	10.5→18.1	12	6.0→8.0
	HiDream	16	22.5→38.8	21.5→37.1	26.1→36.0	26.6→37.6	38.7→62.0	5.6→8.9	5	10.6→19.6
	GLM-Image	65	19.9→38.9	17.5→36.9	21.4→40.5	12.1→25.5	39.8→60.2	5.5→16.3	15	15.1→20.1
	FLUX.1	3	13.4→54.5	16.8→61.8	19.3→54.0	22.3→57.2	35.5→72.1	5.6→22.5	14	25.8→51.1
	mean gain		+16.6	+18.5	+12.0	+15.6	+17.5	+6.9		+9.9
DiT, after norm	FLUX2-klein	246	62.4→63.2	62.2→63.9	62.0→63.0	46.4→47.6	73.0→74.8	22.1→22.4	166	66.5→67.1
	Z-Image	374	58.6→58.4	61.6→61.2	56.8→55.9	41.3→43.5	68.0→68.5	20.1→20.1	101	40.9→40.9
	SD3.5-L	341	52.9→53.3	57.4→57.8	54.9→55.0	47.6→50.1	66.8→68.0	20.9→21.0	28	24.3→30.4
	FLUX.1	289	54.0→54.4	58.9→59.1	53.8→54.3	51.5→57.2	69.1→71.1	22.1→22.4	67	41.2→46.2
	ERNIE-Image	185	48.5→55.4	49.5→58.6	50.5→55.6	42.4→46.7	59.7→65.9	19.7→21.0	48	47.6→53.6
	Krea2	258	46.3→45.8	47.1→46.4	41.4→40.5	32.6→33.6	59.6→57.8	13.4→13.3	93	38.4→38.5
	PixArt-α	162	43.4→45.4	47.5→49.8	48.2→49.6	35.8→38.3	56.7→57.1	18.0→18.2	25	6.1→7.5
	Qwen-Image	309	43.2→43.0	43.5→43.2	39.2→39.0	38.4→39.6	61.5→61.0	13.4→13.3	72	25.4→26.2
	HiDream	174	42.8→43.1	41.9→42.0	40.2→40.4	27.2→31.1	60.4→61.6	8.8→9.1	41	15.7→16.0
	GLM-Image	208	34.7→39.1	33.0→37.3	37.7→40.9	25.9→26.1	55.2→59.2	15.7→16.1	54	13.3→19.0
	mean gain		+1.4	+1.7	+0.9	+2.5	+1.5	+0.3		+2.6

Table 1: Every readout on six of the eleven correspondence evaluations and on k-NN ImageNet classification, shown as cosine → Eq. (3). Teal marks an improvement and red a drop. α is selected per readout on SPair-71k validation and used unchanged everywhere. kNN-IN is top-1 accuracy with $k = 20$ and the same α. It is not among the 506 pairs, and $d_{\text{eff}}^{\text{cls}}$ is computed on the global vector it compares. The other five evaluations are in Appendix D and the full k-NN table in Appendix F.

get worse, all lose 0.2 points or less except MetaCLIP2 (0.7), and all are in the better-conditioned half of the table. One (iBOT) is unchanged. The gain follows the d_{eff} of the global vector that k-NN compares, with $\rho(d_{\text{eff}}, \text{gain}) = -0.80$ and -0.57 after controlling for the baseline, a tighter relation than on any correspondence benchmark. Against the *patch* d_{eff} of the same models it is only -0.35. Choosing k per configuration gives +5.5 instead of +5.2, and choosing α on the test split as well gives +7.2 with 43 of 46 improving. Neither changes which readouts gain most (Table 11).

On the SPair-71k leaderboard (Table 8, Appendix D), diffusion transformers lead under both metrics. After whitening, however, three of the top five places go to raw readouts, compared with one under cosine, and raw FLUX.1 moves from 38th to 3rd.

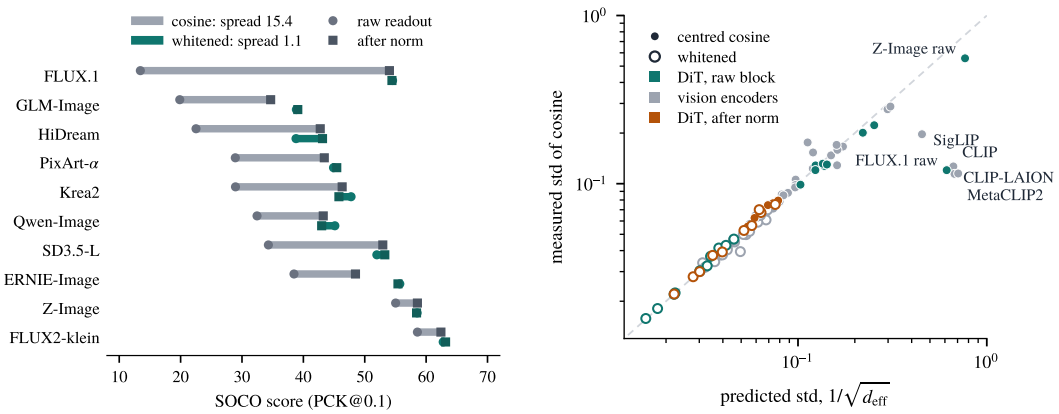

Figure 5: **Left:** ten diffusion transformers, each read at two points with *the same weights*. Each bar spans the two SOCO scores. Under cosine the choice of readout changes the score by 15.4 points on average, and after whitening by 1.1. **Right:** Eq. (2) on real features. The standard deviation of the cosine between patches of different images is plotted against $1/\sqrt{d_{\text{eff}}}$. Filled points are centred cosine. Hollow points are whitened cosine, plotted against the d_{eff} of the whitened covariance. The dashed line is the prediction, with nothing fitted. Labelled readouts have $d_{\text{eff}} < 5$ and spread less than predicted.

The 45 drops are small, with a median of 0.36 points, 38 of them below one point and none above 1.8. They fall into two groups. Twenty-seven are on three post-normalised DiT readouts (Qwen-2512, Krea2 and Z-Image), which have among the highest d_{eff} in the study and little to gain. Sixteen are on SOC-Geo, the geometry-aware variant of SOCO. There the mean gain is only $+1.1$, against $+6.0$ on SOCO, although the rank correlation with d_{eff} (-0.56) is similar to the other benchmarks. The gain is predictable but small, which suggests that cosine was not hiding much geometric information. Two cases are harder to read. CLIP-B and raw FLUX.1 have nearly the same d_{eff} but gain $+4.9$ and $+41.1$, and raw Z-Image, which has the lowest d_{eff} of all, gains only $+3.5$. Section 5.2 returns to both.

5.2 The role of d_{eff}

Diffusion transformers let us vary d_{eff} alone. The same weights can be read before and after the model's own normalisation, so the two readouts carry nearly the same information but differ in how their channels are scaled, and their d_{eff} differs by a median factor of six. If d_{eff} decides how well cosine works, the readout with the lower d_{eff} should score worse under cosine, and whitening should remove the difference. In all ten models the raw readout, which has the lower d_{eff}, scores lower under cosine, by 15.4 SOCO points on average and 40.6 for FLUX.1. After whitening the difference is 1.1 points on average and at most 4.3 (Figure 5, left). The normalisation layer raises d_{eff} by rescaling each channel, and whitening does so from data for any model.

We also tested Eq. (2) directly. For every readout we took 20,000 pairs of patches from different images of the moment corpus and measured their cosine (Figure 5, right). The average is $\|\boldsymbol{\nu}\|^2$, as Eq. (4) predicts, and it is often high (0.87 for raw FLUX.1, 0.99 for raw Z-Image and ERNIE, 0.82 for DINOv3-L). After centring, the standard deviation is within 10% of $1/\sqrt{d_{\text{eff}}}$ for 35 of the 46 readouts and within 25% for 38. After whitening it matches the prediction for all 46, and for 43 within 10%. Six of the eight readouts where the prediction is off are the six with $d_{\text{eff}} < 5$. For the three CLIP variants, SigLIP and raw FLUX.1, the measured spread is two to six times smaller than predicted. Raw Z-Image has almost all its variance in one direction, so its cosines pile up near ± 0.5. The approximation behind Eq. (2) becomes less accurate as d_{eff} gets small (Appendix B.1), so these misses are expected. For such readouts d_{eff} is only a rough summary, and CLIP-B, raw FLUX.1 and raw Z-Image are all among them.

Across all readouts, the rank correlation between d_{eff} and the gain is -0.62 on SOCO (Figure 7), -0.60 on SPair-71k and -0.59 on ScanNet. Low-d_{eff} readouts also score low under cosine and so have more room to gain, but controlling for the cosine score still leaves -0.45.

transform	encoders	DiT raw	DiT norm.	all
LayerNorm	$+0.02$	-0.02	-0.03	$+0.00$
centring, $\mathbf{x} - \boldsymbol{\mu}$	$+1.19$	$+3.30$	$+1.04$	$+1.62$
all-but-the-top (Mu et al., 2018)	$+1.91$	$+9.48$	-0.25	$+3.09$
$\operatorname{diag} \hat{\boldsymbol{\Sigma}}$	$+2.05$	$+16.23$	$+1.01$	$+4.90$
Eq. (3), full	$\mathbf{+3.67}$	$\mathbf{+16.65}$	$\mathbf{+1.43}$	$\mathbf{+6.00}$

Table 2: Parts of the correction. Mean gain over cosine on SOCO for all 46 readouts. Rescaling each channel ($\operatorname{diag} \hat{\boldsymbol{\Sigma}}$) already gives a median 90% of the full gain per readout.

moments from	encoders	DiT raw	DiT norm.	all
corpus (default)	$+3.67$	$+16.65$	$+1.43$	$+6.00$
single image	$+3.41$	$+13.29$	$+0.97$	$+5.03$
image pair	$+2.22$	$+10.12$	$+0.09$	$+3.47$

Table 3: Source of the mean and covariance. Mean gain over cosine on SOCO with moments from the corpus, from each image separately, or from both images pooled. The last two need no data outside the evaluation.

5.3 Ablations

Table 2 compares the full correction with its parts on SOCO. Eq. (4) predicts that centring and whitening fix different things, the mean score and the spread, as the table shows. Centring alone gives $+1.62$ of the full $+6.00$, and the covariance gives the rest. LayerNorm over channels moves 42 of 46 readouts by less than 0.1 points, since it rescales each patch by one factor, which cosine ignores. All-but-the-top (Mu et al., 2018), which removes the few largest principal directions, gives $+3.09$. Using only the diagonal of $\hat{\boldsymbol{\Sigma}}$, which rescales each channel to unit variance and ignores correlations, gives $+4.90$. This is a median 90% of the full gain per readout, and 99% on raw diffusion-transformer readouts. So the main problem is that channels have very different scales, not that they are correlated. The coefficient of variation of the per-channel variances, $\mathrm{CV} = \mathrm{sd}(\operatorname{diag} \boldsymbol{\Sigma}) / \mathrm{mean}(\operatorname{diag} \boldsymbol{\Sigma})$, measures this spread over *channels* rather than directions, and predicts the gain better than d_{eff} ($\rho = +0.64$ against -0.62), as in NLP (Timkey & Van Schijndel, 2021).

Removing correlations still matters in some cases. It gives a median 20% of the gain on encoders and is needed where the diagonal alone fails. On SD2.1 (t=0) the diagonal version *loses* 0.4 points while the full transform gains 4.0, and on CleanDIFT the diagonal gives only 11% of a 4.1-point gain. Both are diffusion U-Nets, whose channels are strongly correlated.

A single image already provides about 10^3 patch samples, and Table 10 (Appendix E) shows that this is nearly enough. Raw FLUX.1 rises from 13.4 to 49.0 with one image and to 52.7 with five, against 54.5 with the default 250. Every other readout that gains more than 0.4 points reaches 90% of its gain within two images. The moments can therefore come from the images being matched, with no data beyond the evaluation. Whitening each image with the moments of its own patches gives a median 84% of the corpus gain and improves 41 of 46 readouts (Table 3). Pooling the patches of the source and target images into one estimate does worse (76%, 35 of 46 improving) and falls below cosine on eleven readouts, for reasons we have not found.

6 Limitations and conclusion

Whitening is not new, and most of its effect comes from rescaling each channel, close to what is already done in NLP. Our contribution is to measure how much it changes zero-shot evaluation, and why. Eq. (8) explains why d_{eff} orders the gains but not their size, because we measured the spread of wrong-candidate scores (Figure 5, right) and not the similarity s of the true match. The k-NN readout uses the layer chosen for correspondence, which may not be where global features are best (Bolya et al., 2026), so those numbers are provisional. Whitening also compresses the spread of the remaining scores (Fune, 2026), likely why the best-conditioned readouts lose a little.

Zero-shot benchmarks of frozen features measure the features and the metric together. How well cosine works depends on d_{eff}, which ranges from 1.7 to 374 across models, so cosine comparisons partly compare feature scaling. Whitening with two moments from unlabeled images removes most of this dependence. It improves 459 of 506 evaluations, shrinks the gap between two readouts of the same weights from 15.4 to 1.1 points, and moves raw FLUX.1 from 38th to 3rd on SPair-71k. We recommend reporting whitened scores next to cosine, and will release our code.

USE OF LARGE LANGUAGE MODELS

We used LLM assistants to aid and polish the writing, to help with plotting code, and to suggest and check some of the analyses. The authors verified all results and take full responsibility for the content of the paper.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=rLX7Vyyzus.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11944–11953, 2022. doi:10.1109/CVPR52688.2022.01164. URL https://ieeexplore.ieee.org/document/9878857/.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Stephane Boucheron, Gabor Lugosi, and Pierre Massart. Concentration inequalities: A non asymptotic theory of independence, 2013.

Hilton Bristow, Jack Valmadre, and Simon Lucey. Dense semantic correspondence where every pixel is a classifier. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4022–4030, 2015. doi:10.1109/ICCV.2015.458. URL https://ieeexplore.ieee.org/document/7410815/.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International conference on learning representations*, 2021.

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Chanwoo Chun, Abdulkadir Canatar, SueYeon Chung, and Daniel Lee. Estimating dimensionality of neural representations from finite samples. In *International Conference on Learning Representations*, volume 2026, pp. 93949–93979, 2026.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H196sainb.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Olaf Dünkel, Basavaraj Sunagad, Haoran Wang, David T Hoffmann, Christian Theobalt, and Adam Kortylewski. Soco: Benchmarking semantic object correspondence in vision foundation models. In *European Conference on Computer Vision*, pp. 251–269. Springer, 2026.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Jerome H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. doi:10.1080/01621459.1989.10478752. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478752.

Ernesto Lopez Fune. High-dimensional concentration and retrieval instability in embedding spaces, 2026. URL https://arxiv.org/abs/2606.28330.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=s3MwCBuqav.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, pp. 214262, 2017.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Is anisotropy inherent to transformers? *arXiv preprint arXiv:2306.07656*, 2023.

Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision – ECCV 2012. 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, volume 7575 of *Lecture Notes in Computer Science*, pp. 459–472. Springer, 2012. doi:10.1007/978-3-642-33765-9_33. URL https://link.springer.com/chapter/10.1007/978-3-642-33765-9_33.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22487–22497, 2025.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. URL https://navidataset.github.io/.

Hervé Jégou and Ondrej Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *Computer Vision – ECCV 2012. 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, pp. 774–787. Springer, 2012. doi:10.1007/978-3-642-33709-3_55. URL https://link.springer.com/chapter/10.1007/978-3-642-33709-3_55.

Euna Jung, Jungwon Park, Jaekeol Choi, Sungyoon Kim, and Wonjong Rhee. Isotropic representation can improve dense retrieval. In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part III*, pp. 125–137, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-33379-8. doi:10.1007/978-3-031-33380-4_10. URL https://doi.org/10.1007/978-3-031-33380-4_10.

András Kalapos and Bálint Gyires-Tóth. Whitening consistently improves self-supervised learning. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 448–453, 2024. doi:10.1109/ICMLA61862.2024.00066.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is CLIP ideal? No. Can we fix it? Yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22436–22446, 2025. doi:10.1109/ICCV51701.2025.02083. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Kang_Is_CLIP_ideal_No._Can_we_fix_it_Yes_ICCV_2025_paper.pdf.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. Bert busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3392–3405, 2021.

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

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi:10.1016/S0047-259X(03)00096-4. URL https://doi.org.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, pp. 671–689. Springer, 2026.

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

V. S. Raghu Parupudi. Anisotropy decides cosine vs. rank metrics for text embeddings. *arXiv preprint arXiv:2606.29571*, 2026.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018. doi:10.1109/TPAMI.2018.2846566. URL https://ieee.org.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Radovan Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86): 2487–2531, 2010. URL http://jmlr.org.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3325–3339, 2022.

Mert Búlent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30084–30094, June 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research (TMLR)*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, pp. 887–890, 2024.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 117–127. IEEE, 2025.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

Mingjie Sun, Mingi Chen, Zichao Liu, and Tiejun Huang. Massive activations in large language models. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17923–17934. IEEE, 2023.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/903ce9225fca3e988c2af215d4e544d3-Paper-round2.pdf.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3076–3085. IEEE, 2024.

Kaiyan Zhang, Xinghui Li, Jingyi Lu, and Kai Han. Semantic correspondence: Unified benchmarking and a strong baseline. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

A MODELS AND CONFIGURATIONS

Table 4 gives the diffusion-transformer readouts and Table 5 the vision encoders.

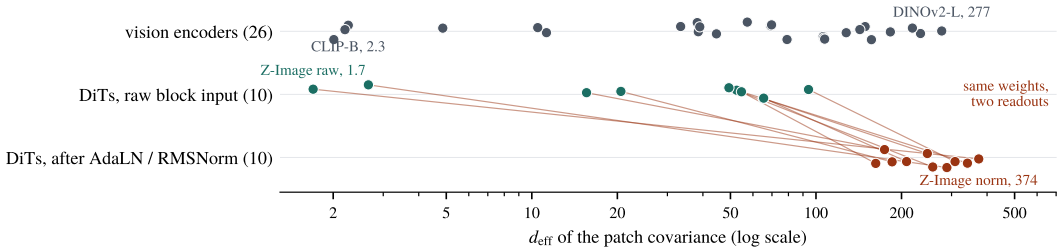

Figure 6: Effective dimension of the patch covariance for all 46 readouts, by family. The ten diffusion transformers appear twice, read at a block's input and after the block's own normalisation, and a line joins the two readouts of each; the same weights differ by a median factor of six.

model	block	timestep	d	patch	massive-activation channels
ERNIE-Image	7	600	4096	16	317, 382, 942, 1180
FLUX2-klein	11	750	4096	16	1395, 1868, 2392
FLUX.1	28	261	3072	16	154, 1446
GLM-Image	12	550	4096	16	2430, 2964, 3399, 3731
HiDream-O1	20	300	4096	32	1631, 1838, 2276, 2391
Krea2-turbo	12	450	6144	16	3389, 5895
PixArt-α-512	14	141	1152	16	293
Qwen-Image-2512	37	375	3072	16	673
SD3.5-Large	23	380	2432	16	676
Z-Image-turbo	21	425	3840	16	85, 1238

Table 4: Extraction settings for the ten diffusion transformers. For the seven models without published settings, block and timestep are chosen on the SPair-71k validation split, maximising the cosine score, and fixed for every benchmark. One noise sample per image. The published settings of Gan et al. (2025) are used where they exist (FLUX.1, SD3.5-Large, PixArt-α); the other seven have not previously been evaluated as dense descriptors. Massive-activation channels are the coordinates carrying anomalous variance at the raw block input, listed for reference against Gan et al. (2025); whitening itself uses no knowledge of them, but the post-normalisation readouts drop them as DiTF does. HiDream-O1 is the only model with a 32-pixel patch, so it yields a quarter as many patch vectors per image as the rest.

Scoring criteria. Following the SOCO protocol (Dünkel et al., 2026), the argmax of (1) is taken over the raw patch grid of the target image and the winning patch centre is the prediction; features are not upsampled to pixel resolution, so K is the patch count, 10^3 to 10^4 at the resolutions used here: SOCO, SPair-71k and AP-10K at 800 px (rounded to the patch size), ScanNet and NAVI at a quarter of native resolution, TAP-Vid at 448 px, k-NN at 224 px. The threshold differs by task and we use each benchmark's own: for SOCO, CC, Cross-SOC, SPair-71k and AP-10K, recall at 0.1 of the bounding-box scale, averaged per image and then over images, which the SPair-71k literature reports as PCK@0.1; for ScanNet, recall within 10 absolute pixels of the reprojected correspondence; for NAVI, recall within 2 cm in 3D restricted to pairs whose relative camera rotation lies in the 30–60° bin, as in the benchmark paper; and for TAP-Vid, Average Jaccard, which scores position and occlusion prediction jointly.

B DERIVATIONS

Throughout, $\mathbf{x}, \mathbf{y}$ are independent draws from a distribution on $\mathbb{R}^d$ with covariance $\mathbf{\Sigma}$, eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$, and $d_{\text{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$.

encoder	reference	d	d_{eff}	$\lambda_1/\operatorname{Tr}\boldsymbol{\Sigma}$	CV
C-RADIO-v3-B	Heinrich et al. (2025)	768	157	0.034	0.7
CLIP-B/16	Radford et al. (2021)	768	2	0.664	16.3
CLIP-B/16 (LAION)	Cherti et al. (2023)	768	2	0.674	17.4
CleanSD2.1 (t=0)	Stracke et al. (2025)	1280	106	0.046	1.2
ConvNeXt-B (IN22k)	Liu et al. (2022)	1024	79	0.099	2.6
CroCo v2	Weinzaepfel et al. (2023)	768	11	0.232	7.9
SD2.1 (t=0)	Tang et al. (2023)	1280	107	0.042	1.2
SD2.1 (t=250)	Tang et al. (2023)	1280	149	0.039	0.9
DINO-B/16	Caron et al. (2021)	768	107	0.076	1.0
DINOv2-B/14	Oquab et al. (2024)	768	233	0.029	0.7
DINOv2-L/14	Oquab et al. (2024)	1024	277	0.026	0.7
DINOv3-B/16	Siméoni et al. (2026)	768	128	0.045	1.0
DINOv3-L/16	Siméoni et al. (2026)	1024	143	0.039	1.0
DUNE-B/14	Sarıyıldız et al. (2025)	768	33	0.153	2.4
DUNE-B/14 (448)	Sarıyıldız et al. (2025)	768	57	0.102	1.9
DeiT III-B/16	Touvron et al. (2022)	768	39	0.122	3.7
I-JEPA-H/16 (448)	Assran et al. (2023)	1280	69	0.060	3.0
MAE-B/16	He et al. (2022)	768	45	0.102	1.3
MetaCLIP2-B/16	Chuang et al. (2026)	768	2	0.705	18.3
MiDaS-L/16	Ranftl et al. (2020)	1024	218	0.031	0.3
Perception Enc.-B/16 (512)	Bolya et al. (2026)	768	38	0.124	0.5
SAM-B	Kirillov et al. (2023)	768	39	0.108	1.9
SigLIP-B/16	Zhai et al. (2023)	768	5	0.422	7.5
V-JEPA 2.1-B	Mur-Labadia et al. (2026)	768	70	0.075	2.0
VGGT	Wang et al. (2025)	2048	10	0.233	4.9
iBOT-B/16	Zhou et al. (2022)	768	183	0.031	0.6

Table 5: The 26 vision encoders. d is the feature width, d_{eff} the participation ratio of $\boldsymbol{\Sigma}$, $\lambda_1/\operatorname{Tr}\boldsymbol{\Sigma}$ the share of variance in the leading direction, and CV the coefficient of variation of the per-channel variances. All statistics are computed on the corpus of Section 5.

B.1 The null score

Write $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$ for the direction of a patch feature, so that $\cos(\mathbf{x}, \mathbf{y}) = \langle \mathbf{u}_\mathbf{x}, \mathbf{u}_\mathbf{y} \rangle$ with the two independent, and let $\boldsymbol{\nu} = \mathbb{E}[\mathbf{u}]$ and $R = \mathbb{E}[\mathbf{u}\mathbf{u}^\top]$, both of which exist whenever second moments do. Then

$$\mathbb{E}\big[\cos(\mathbf{x}, \mathbf{y})\big] = \|\boldsymbol{\nu}\|^2, \qquad \operatorname{Var}\big(\cos(\mathbf{x}, \mathbf{y})\big) = \operatorname{Tr}(R^2) - \|\boldsymbol{\nu}\|^4. \tag{4}$$

Both are exact and distribution-free. Two patches that share nothing therefore score $\|\boldsymbol{\nu}\|^2$ on average, and this offset is a property of the feature distribution alone. A cloud that sits far from the origin has small angles between all of its members whatever they encode.

Centring removes the first term. If the distribution is centrally symmetric about the origin then $\boldsymbol{\nu} = 0$, the offset vanishes, and $\operatorname{Var}(\cos) = \operatorname{Tr}(R^2)$. Expanding $\|\mathbf{x}\|^2$ about $\operatorname{Tr}\boldsymbol{\Sigma}$ gives $R = \boldsymbol{\Sigma}/\operatorname{Tr}\boldsymbol{\Sigma} + O(1/d_{\text{eff}})$ and so (2), $\operatorname{Var}(\cos) \approx 1/d_{\text{eff}}$, with relative error $O(1/d_{\text{eff}})$, tightest where d_{eff} is large and loosest at the concentrated end, where the $[-1, 1]$ bound also binds.

Equation (4) therefore separates the two failures the transform of (3) repairs, and assigns one operation to each. Subtracting $\boldsymbol{\mu}$ brings the mean null score $\|\boldsymbol{\nu}\|^2$ close to zero (exactly zero for a centrally symmetric distribution), whatever the conditioning. Whitening raises d_{eff} towards d, which narrows what remains. Neither substitutes for the other: a centred but badly conditioned readout still has a wide null, and a well-conditioned but off-centre one still has a raised floor. Section 5.3 measures the two contributions separately. Because the baseline in (1) applies cosine without centring, its null carries the offset as well, and (2) alone understates the difficulty of the comparison it faces.

B.2 Estimating d_{eff}

The participation ratio is biased at finite samples (Chun et al., 2026), and Section 3 quotes it to three figures, so the direction and size of that bias matter. For a sample covariance $\hat{\boldsymbol{\Sigma}}$ from n independent draws the Wishart moments give $\mathbb{E}[\operatorname{Tr}\hat{\boldsymbol{\Sigma}}^2] = \frac{n+1}{n}\operatorname{Tr}(\boldsymbol{\Sigma}^2) + \frac{1}{n}(\operatorname{Tr}\boldsymbol{\Sigma})^2$ and $\mathbb{E}[(\operatorname{Tr}\hat{\boldsymbol{\Sigma}})^2] = (\operatorname{Tr}\boldsymbol{\Sigma})^2 +$

$\frac{2}{n}\mathrm{Tr}(\boldsymbol{\Sigma}^2)$, so to first order

$$\hat{d}_{\mathrm{eff}} \;\approx\; d_{\mathrm{eff}}\Big(1-\frac{d_{\mathrm{eff}}}{n}\Big). \tag{5}$$

The bias is downward and grows with d_{eff}, so it compresses the range we report rather than inflating it. With the corpus of Section 5 the largest relative bias over the 46 readouts is 0.11%, on HiDream-O1 after normalisation, which combines the fewest patch vectors (1.6×10^5, since it patchifies at 32) with a high d_{eff}; no other readout exceeds 0.06%. Patches within an image are correlated, so the effective n is smaller than the raw count, but the bound stays under 1.2% even if correlation costs a factor of ten.

B.3 The bar a match must clear

A query is answered correctly only if the true match outscores all $K-1$ distractors, so what a match must beat is not a typical null score but the largest of many. Write $Z_j=\cos(\mathbf{x},\mathbf{y}_j)$ for the score against distractor j, taken after centring so that $\mathbb{E}Z_j=0$ (Appendix B.1); without centring every Z_j is raised by the same offset $\|\boldsymbol{\nu}\|^2$ and so is the bar. We take the Z_j to be independent and sub-Gaussian with parameter $\sigma=d_{\mathrm{eff}}^{-1/2}$, that is $\mathbb{E}\,e^{tZ_j}\le e^{t^2\sigma^2/2}$ for all t. The parameter is the one (2) supplies: $\mathrm{Var}(Z_j)\approx 1/d_{\mathrm{eff}}$. Matching parameter to standard deviation is an assumption rather than a consequence, since a variance bounds no tail on its own, but it is approximately right for Gaussian features with large d_{eff}, and cosine's own bound $|Z_j|\le 1$ makes the Z_j sub-Gaussian regardless, with a parameter that is merely looser.

Under that assumption the maximum obeys the standard bound. For any $t>0$, Jensen applied to $\exp$ gives

$$\mathbb{E}\Big[\max_{j<K} Z_j\Big] \;\le\; \frac{1}{t}\,\ln\sum_{j<K}\mathbb{E}\,e^{tZ_j} \;\le\; \frac{\ln K}{t}+\frac{t\sigma^2}{2}, \tag{6}$$

and $t=\sigma^{-1}\sqrt{2\ln K}$ minimises the right-hand side, so

$$\mathbb{E}\Big[\max_{j<K}\cos(\mathbf{x},\mathbf{y}_j)\Big] \;\lesssim\; \sqrt{\frac{2\ln K}{d_{\mathrm{eff}}}}, \tag{7}$$

and the bound is asymptotically tight as $K\to\infty$ for independent Gaussian Z_j (Boucheron et al., 2013). The bar therefore grows with the logarithm of the candidate count and falls with $\sqrt{d_{\mathrm{eff}}}$: increasing K from 10^4 to 10^6 raises the bar by a fifth, the same as giving up a third of d_{eff}.

Two remarks on when (7) says anything. Cosine is bounded by 1, so the bound is vacuous once its right-hand side reaches that value, which it does at $K=10^4$ for every readout with $d_{\mathrm{eff}}<18$. In that regime the distractors are pinned against the ceiling rather than spread below it, and the contest is settled by whatever noise separates scores that are all close to 1; Section 3 reads off exactly this for the raw readouts. Independence is also an idealisation. In (1) the candidates come from a single target image and are correlated with each other and with the query. Positive correlation among the distractors lowers their expected maximum, while correlation with the query can raise it, so (7) describes how the bar scales with K and d_{eff} rather than its exact value.

B.4 Accuracy, and the gain from whitening

Let s denote the similarity at the true match, itself random across queries. Under the same independence and Gaussian approximation, with Φ the standard normal c.d.f.,

$$\mathbb{P}(\text{correct}) \;=\; \mathbb{E}_s\Big[\Phi\big(s\sqrt{d_{\mathrm{eff}}}\big)^{K-1}\Big]. \tag{8}$$

Three consequences follow. The results of Section 5.2 are consistent with the first two, and we do not test the third.

(i) Saturation in d_{eff}. The integrand is a sigmoid in $s\sqrt{d_{\mathrm{eff}}}$, so raising d_{eff} helps a great deal while $s\sqrt{d_{\mathrm{eff}}}$ is small and almost not at all once it is large. The relation between d_{eff} and accuracy is therefore monotone and saturating rather than linear, which is why a rank correlation across readouts is the right summary and why it cannot be expected to be tight.

(ii) Saturation in the baseline. A model whose s is already large sits on the plateau. Gains and baseline scores are therefore mechanically anti-correlated, independently of d_{eff}, which is what a partial correlation controlling for the baseline removes.

(iii) Dependence on K. The exponent is $K - 1$, so accuracy falls and the available gain rises with $\ln K$; a resolution sweep, which varies K directly, would test this.

B.5 Reparameterisation

For invertible $\boldsymbol{A}$ the features $\boldsymbol{A}\mathbf{x}$ carry the same information and have covariance $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$. Taking $\boldsymbol{A} = \operatorname{diag}(a_1, \ldots, a_d)$ in the eigenbasis of $\boldsymbol{\Sigma}$ scales $\lambda_i \mapsto a_i^2 \lambda_i$, and d_{eff} is continuous in $(a_1^2\lambda_1, \ldots, a_d^2\lambda_d)$, equal to d when all are equal and tending to 1 as one dominates. Any value in $(1, d]$ is therefore attained.

C Sensitivity to how α is chosen

Table 7 shows what the transfer protocol costs on every evaluation. Table 6 scores five rules for setting α on the four evaluations with a full sweep (SPair-71k validation, ScanNet, NAVI, TAP-Vid; 184 readout–benchmark cells). The reported protocol selects α per readout on the SPair-71k validation split and transfers it; the oracle picks the best value per cell with the test labels in hand and is an upper bound no protocol can reach. Every rule gives the same three families in the same order, and the spread between choosing α with no information at all ($\alpha = 0$) and with full knowledge of the test scores is one point on an effect of $+5.7$. The choice of α changes the size of the gain but not its sign. The val-selected rule is not the best mean but has the fewest failures and by far the smallest worst case, because it alone lets the well-conditioned post-normalised readouts choose $\alpha \to 1$ and be left nearly alone, as Section 3 says they should be.

rule	DiT raw	encoders	DiT norm.	all	improving	worst
oracle per cell (upper bound)	+15.38	+4.84	+2.86	+6.70	97.3%	–
selected on SPair-71k val	+14.42	+4.04	+1.48	+5.74	94.6%	−1.75
fixed $\alpha = 0.1$	+14.67	+4.30	+1.98	+6.05	88.0%	−7.16
fixed $\alpha = 0.5$	+14.06	+3.82	+2.10	+5.67	93.5%	−6.42
$\alpha = 0$, no choice at all	+14.37	+3.96	+1.37	+5.66	85.9%	−8.75

Table 6: Mean gain over cosine under five rules for α, pooled over the four evaluations with a full sweep. The reported rule is in bold.

evaluation	reported (α from val)	per-cell oracle	share
SOC	+6.00	+6.05	99%
CC	+7.29	+7.35	99%
Cross-SOC	+6.41	+6.48	99%
SPair-71k	+6.93	+6.97	99%
AP-10K	+4.86	+4.94	98%
AP-10K-C.S.	+5.60	+5.66	99%
AP-10K-C.F.	+4.82	+4.91	98%
SOC-Geo	+1.05	+1.23	86%
ScanNet	+7.58	+10.04	76%
NAVI	+6.43	+7.61	84%
TAP-Vid	+2.25	+2.41	93%
all eleven	+5.38	+5.79	93%

Table 7: What the protocol of Section 5 costs. The reported column selects α once per readout on SPair-71k validation and transfers it; the oracle picks the best α for each readout and evaluation using the test labels, and is not attainable. Transfer is nearly free within the semantic group and costs most on ScanNet and NAVI, whose oracle α sits near 0 while SPair's sits near 0.9.

	readout	cosine	corrected	rank under cosine
1	FLUX2-klein, after norm	62.2	**63.9**	1
2	FLUX2-klein, raw	54.5	**62.9**	10
3	FLUX.1, raw	16.8	**61.8**	38
4	Z-Image, raw	58.0	**61.5**	4
5	Z-Image, after norm	61.6	**61.2**	2
6	DINOv3-L/16	56.5	**60.0**	7
7	FLUX.1, after norm	58.9	**59.1**	3
8	DINOv3-B/16	55.7	**58.8**	9
9	ERNIE, raw	38.2	**58.6**	25
10	ERNIE, after norm	49.5	**58.6**	13

Table 8: SPair-71k, PCK@0.1, corrected leaderboard (top 10 of 46). Under cosine the top five are FLUX2-klein after norm, Z-Image after norm, FLUX.1 after norm, Z-Image raw and SD3.5-L after norm; the correction brings raw readouts up to their post-normalised counterparts rather than changing which family leads. All runs are at 800 px, PCK averaged per image.

Sweeping α on each benchmark's test split (done only to measure what the validation protocol of Section 5 gives up, never to set a reported value) shows it to be a property of the task family more than of the model. On ScanNet 30 of 46 readouts prefer $\alpha = 0$, full whitening; on NAVI the mode is 0.1; on SOCO it is 0.9, mild shrinkage, and the other semantic benchmarks mostly peak there too (SOC-Geo at 1.0). TAP-Vid is bimodal, 12 readouts at 0.0 and 11 at 0.9, and so does not sort with the geometric tasks it resembles. Table 6 scores every selection rule from $\alpha = 0$ to the per-cell oracle, and the spread between them is one point on a $+5.7$ effect.

A closed-form α. Ledoit–Wolf shrinkage (Ledoit & Wolf, 2004) chooses α to minimise the expected error of $\hat{\mathbf{\Sigma}}$ as an estimate of $\mathbf{\Sigma}$, toward the same target $\bar{\sigma}^2\mathbf{I}$. On $50{,}000$ patches from the moment corpus it returns $\alpha_{\mathrm{LW}} < 0.01$ for all 46 readouts (median 0.003), where the sweep returns values from 0 to 1. Even $50{,}000$ patches, 8 to 65 per dimension, estimate the covariance well enough that shrinkage has little to repair. The α of (3) is therefore not an estimation device. It sets how far toward full whitening the metric goes, and the sweep says the semantic benchmarks prefer to stop short ($\alpha \approx 0.9$) while ScanNet and NAVI prefer to go all the way. Only when the corpus is a single image, with fewer patches than dimensions, does shrinkage become necessary, and Table 10 shows the transferred α suffices there.

D Every evaluation

Table 9 gives the five evaluations not in Table 1 and Table 8 the SPair-71k leaderboard. Figure 7 plots the gain against d_{eff}, Figure 8 shows every gain, Figure 9 the change in rank, and Figure 10 two more keypoints of the kind shown in Figure 1.

Figure 8 makes the block structure visible. The raw diffusion-transformer readouts form a solid band and the failures sit in two corners.

E How much data the moments need

Which images. To check that the gains do not depend on the particular 250 images, we redrew the corpus from the ImageNet training set with different random seeds, two to five draws per readout, for 24 of the 26 vision encoders. The SOCO gain changes little. Its standard deviation across draws is a median 0.03 points and at most 0.17 (SAM-B), and no readout's gain varies by more than 0.4 points between draws.

Whitening twice. Whitening features that are already whitened, with moments re-estimated on them, changes SOCO by at most 0.4 points on the five readouts we tried (raw FLUX.1 $54.5 \to 54.5$, FLUX.1 after AdaLN $54.4 \to 54.6$, DINOv2-L $62.6 \to 62.6$, CLIP-B $21.1 \to 21.0$, raw SD3.5-L $52.0 \to 51.6$).

	readout	CC	Cross-SOC	AP-10K-C.S.	AP-10K-C.F.	SOC-Geo
vision encoders	DINOv2-L/14	80.4→80.8	56.8→58.1	60.9→61.6	48.8→50.3	72.3→72.3
	DINOv3-L/16	76.7→78.1	52.8→57.9	58.2→60.7	44.9→51.2	74.1→73.7
	DINOv2-B/14	78.8→79.2	54.9→56.2	60.5→61.1	49.3→50.5	71.9→71.7
	DINOv3-B/16	73.3→74.6	50.9→54.7	57.1→59.6	44.8→49.2	74.7→74.4
	C-RADIOv3-B	69.0→69.5	46.3→46.9	49.3→49.4	41.5→41.8	68.9→68.8
	CleanDIFT	62.1→67.7	44.7→48.5	46.7→50.3	32.9→35.6	77.6→76.2
	SD2.1 t=250	60.5→65.4	42.7→45.5	45.5→48.2	31.9→33.4	76.8→75.1
	DUNE-B/14	61.2→65.7	39.4→42.6	44.2→47.7	32.6→36.5	73.5→73.0
	I-JEPA-H/16	60.5→65.9	38.4→43.8	44.1→48.9	31.5→37.4	71.7→71.4
	PE-Spatial-B	63.2→65.4	40.9→42.5	43.5→45.1	34.7→35.9	68.5→68.6
	DUNE-B/14_{448}	58.7→64.1	37.8→41.7	42.5→46.4	32.1→36.1	73.5→73.0
	SD2.1 (t=0)	56.0→62.0	38.3→42.2	43.4→48.1	31.4→34.8	76.3→74.7
	V-JEPA2.1-B	60.4→64.2	38.6→42.3	42.8→46.2	32.6→36.1	69.5→70.0
	iBOT-B/16	55.2→56.4	34.1→35.6	42.2→43.2	32.1→33.8	69.4→69.5
	DINO-B/16	43.8→46.4	23.9→26.0	27.2→29.9	17.6→20.5	68.4→68.5
	MiDaS-L/16	40.5→41.3	21.4→22.5	24.0→25.6	17.5→18.7	66.9→67.1
	SAM-B	35.2→39.0	20.4→22.1	21.0→24.0	14.3→15.9	70.7→71.0
	ConvNeXt-B	34.2→48.5	10.4→21.2	12.1→23.8	5.4→10.8	66.5→67.3
	DeiT3-B/16	35.5→40.4	12.1→17.1	10.8→16.0	5.8→9.2	65.2→65.5
	MetaCLIP2-B	31.2→34.7	15.4→17.8	18.1→20.5	12.8→14.4	65.1→65.4
	CLIP-B/16_{LAION}	26.3→30.7	11.7→15.5	15.4→19.0	10.2→14.0	64.9→65.2
	CLIP-B/16	25.2→32.1	11.4→16.3	15.6→21.2	10.2→14.6	64.7→65.3
	VGGT-1B	21.6→53.1	12.5→33.3	12.0→36.4	9.2→25.7	72.2→73.6
	CroCo-v2	15.2→20.4	7.8→10.6	6.7→8.7	5.5→6.4	67.2→68.7
	MAE-B/16	14.4→18.9	7.2→9.6	7.3→9.3	5.7→7.2	66.5→68.1
	SigLIP-B/16	14.1→17.0	7.0→8.8	6.9→8.9	5.9→7.2	64.5→66.5
	mean gain	+4.9	+3.9	+3.9	+3.3	+0.1
DiT, raw block	FLUX2-klein	76.1→81.2	48.5→58.9	56.2→61.8	38.6→50.4	71.5→72.3
	Z-Image	68.7→72.0	47.5→52.4	52.1→54.3	36.0→39.7	73.5→75.7
	ERNIE-Image	50.5→71.3	29.4→50.5	36.7→53.8	23.5→41.3	70.7→73.7
	SD3.5-L	45.5→65.8	28.2→47.0	35.4→50.2	25.5→37.3	70.2→74.3
	Qwen-Image	42.9→54.9	25.5→38.7	30.7→40.4	20.9→28.4	68.9→74.2
	Krea2	40.6→59.9	22.3→43.8	30.4→40.9	18.2→31.2	66.1→73.2
	PixArt-α	39.6→58.3	23.1→39.9	31.4→46.8	19.6→33.6	68.5→73.0
	HiDream	31.0→48.3	17.1→31.7	20.3→31.1	12.7→17.9	63.6→70.1
	GLM-Image	29.0→51.2	16.5→32.2	18.7→36.2	13.8→25.6	67.6→72.5
	FLUX.1	20.4→69.0	9.7→46.7	14.4→52.2	9.4→35.0	65.6→73.4
	mean gain	+18.8	+17.4	+14.2	+12.2	+4.6
DiT, after norm	FLUX2-klein	80.6→81.5	57.8→59.7	60.6→61.9	48.7→51.0	72.7→72.6
	Z-Image	71.9→71.7	52.4→52.2	54.1→53.0	39.7→37.9	75.7→75.7
	SD3.5-L	66.4→67.1	48.1→48.7	52.3→52.4	40.6→40.9	75.1→74.7
	FLUX.1	68.5→68.7	47.4→47.3	51.0→51.5	36.6→35.5	73.4→73.5
	ERNIE-Image	61.5→71.2	39.9→50.7	46.0→54.0	31.5→42.0	73.5→73.7
	Krea2	58.3→57.5	41.3→40.8	39.6→38.8	29.6→28.8	72.7→72.7
	PixArt-α	55.7→58.8	37.9→40.6	45.1→47.5	32.0→34.6	73.7→73.4
	Qwen-Image	51.4→51.0	36.9→36.6	36.1→36.0	24.9→24.5	75.0→75.0
	HiDream	53.9→54.3	37.6→38.0	36.7→37.4	25.6→25.9	70.1→70.4
	GLM-Image	45.4→51.5	28.3→32.4	33.7→36.8	23.6→26.2	72.6→72.4
	mean gain	+2.1	+2.1	+1.4	+1.5	−0.1

Table 9: The five evaluations omitted from Table 1, same convention. C.S. and C.F. are AP-10K across species and across families. The four semantic splits track the columns of Table 1 closely. SOC-Geo does not. It is the only evaluation on which the three families score nearly alike, and on which post-normalised readouts lose on average.

F k-NN CLASSIFICATION

Following Dünkel et al. (2026), embeddings are extracted over the ImageNet training set and evaluated on the validation set. Features are L_2-normalised and neighbours are scored by cosine similarity with temperature-weighted voting, as in DINO (Caron et al., 2021); both columns of the table go through the same search. Classification uses the CLS token where a model has one and the mean of the dense tokens otherwise. Where Dünkel et al. (2026) select k for best accuracy, we fix $k = 20$: selecting it per configuration would choose a hyper-parameter on the split being reported, which is the practice Section 5 avoids for α. α itself is the SPair-71k validation value of Table 1, transferred unchanged; where that value was not in the k-NN sweep grid the nearest grid point is used (marked †). The last two columns give the oracle numbers anyway, α chosen on the test split and then k as well, and they support the same conclusion. Krea2 and FLUX.1 were indexed on a 100-image-

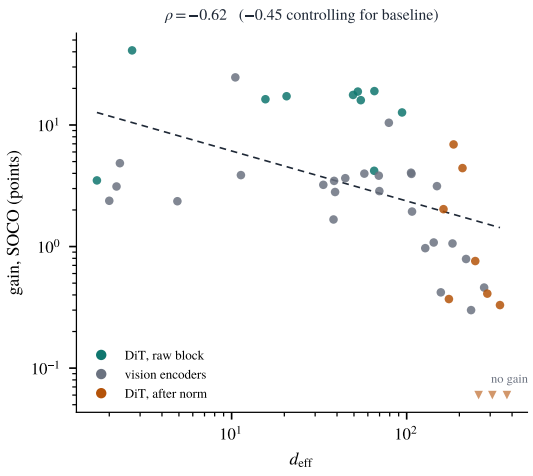

Figure 7: Gain over cosine against d_{eff} on SOCO, on log axes, all 46 readouts; $\rho = -0.62$, and -0.45 controlling for the baseline score. Readouts that gain nothing or lose are marked on the lower edge.

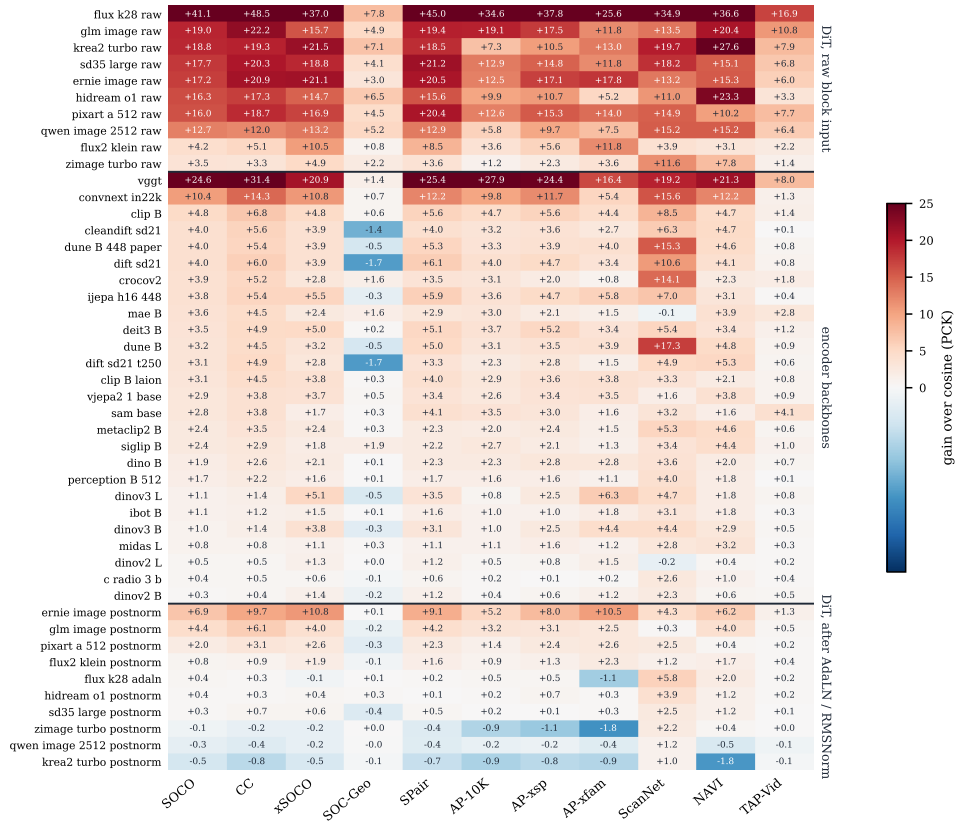

Figure 8: Gain over cosine for all 46 readouts on all eleven benchmarks, grouped by family and sorted within family by SOCO gain.

per-class subset of the training set. The layer is the one selected for correspondence and was not searched again: Bolya et al. (2026) report that a model's best global and best local features need not come from the same layer, so a better layer for classification may exist for several of these readouts.

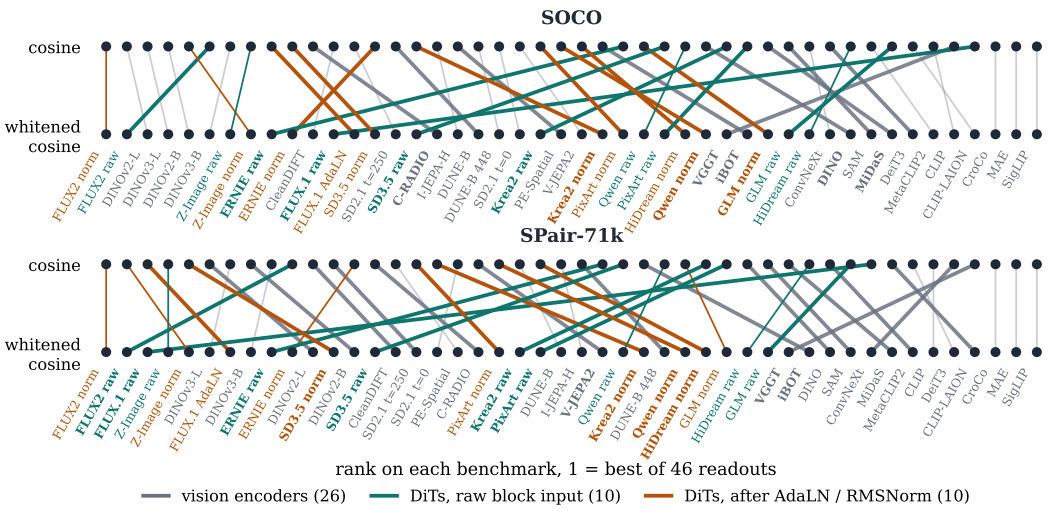

Figure 9: All 46 readouts ranked under cosine and under Eq. (3) on two benchmarks, named at their corrected rank and coloured by family; moves of six ranks or more are bold. Nine of the ten raw DiT readouts rise on both, FLUX.1 by 31 and 35 ranks. Rank is zero-sum, so the encoders that fall here did not lose score (Table 1); on SPair-71k thirteen encoders and five post-normalised readouts drop four ranks or more because raw DiT readouts overtook them.



Figure 10: Two more of the 20 keypoints behind Figure 1, same convention: a sheep matched into a lamb in a different pose, and a person matched across two indoor scenes. The person pair has the largest whitened margin of the twenty and the bus of Figure 1 the second; we lead with the bus because the person keypoint's tolerance is 31 px and the marks would overlap at print size. On every one of the twenty, the four raw DiT readouts miss under cosine and hit after whitening, and both encoders hit under both.

The absolute accuracies are therefore a lower bound on what each backbone can do; what the table measures is cosine against Eq. (3) at a fixed readout, which is unaffected.

G THE MECHANISM IN ISOLATION

Figure 11 shows the mechanism on real features.

		images used to estimate $\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}$									
readout	cosine	1	2	5	10	25	50	100	250	500	n_{90}
FLUX.1, raw	13.4	49.0	48.2	52.7	53.8	54.6	54.6	54.8	54.5	54.4	5
HiDream, raw	22.5	38.4	39.1	37.3	38.0	38.4	38.5	38.7	38.8	39.2	1
Z-Image, raw	55.0	58.4	58.4	58.2	58.1	58.5	58.5	58.6	58.5	58.5	1
CLIP-B/16$_{\text{LAION}}$	16.7	18.8	19.6	19.7	19.8	19.8	19.8	19.8	19.8	19.8	2
FLUX.1, AdaLN	52.9	54.7	54.3	54.3	54.3	54.6	54.7	54.7	54.5	54.5	1
HiDream, norm.	42.8	42.8	43.0	43.0	43.1	43.1	43.1	43.1	43.1	43.1	50
DINOv2-B/14	60.5	60.4	60.6	60.8	60.8	60.8	60.8	60.8	60.8	60.8	100
Z-Image, norm.	58.6	58.5	58.4	58.4	58.5	58.3	58.4	58.4	58.4	58.4	–
Krea2, norm.	46.3	45.8	45.6	46.0	46.0	45.7	45.9	45.8	45.8	45.8	–

Table 10: SOCO score against the number of unlabeled ImageNet images used to estimate the two moments; n_{90} is the number at which 90% of the converged gain is reached, and 250 is the setting used throughout the paper. Every readout that gains reaches 90% of it within five images, all but raw FLUX.1 within two; the two with n_{90} of 50 and 100 gain under 0.4 points, and the two with nothing to gain are flat at every corpus size.

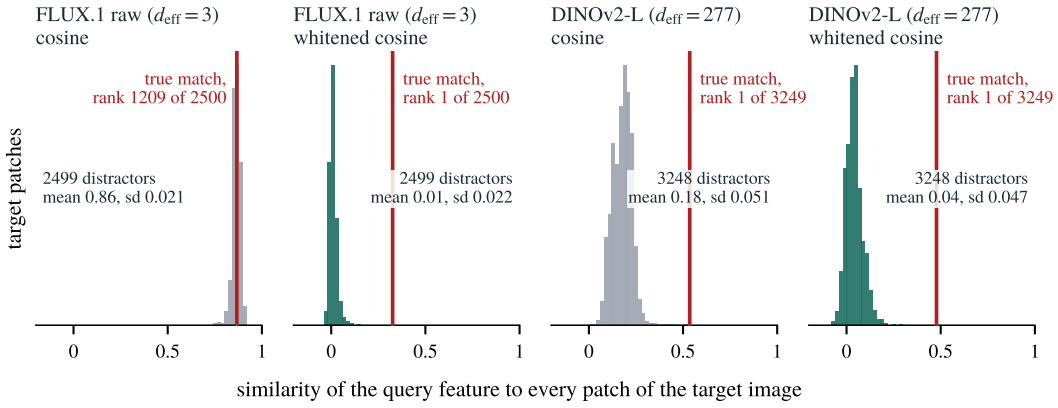

Figure 11: The mechanism on real features. One query patch scored against every patch of the target image (the pair of Figure 1), for a badly and a well-conditioned readout, under cosine and after whitening. For FLUX.1 read raw, cosine puts the $2{,}499$ distractors at 0.86 ± 0.02 and the true match at 0.87, where it ranks 1,209th. Whitening removes the offset and the true match stands alone. DINOv2-L's distractors were already well below its true match under cosine, and whitening changes little.

readout	family	d_{eff}	CV	top-1 (k=20, α transferred)	α	top-1 (k=20, oracle α)	top-1 (best k, oracle α)
FLUX.1, raw	DiT	14	5.3	25.8→51.1	0.0	25.8→51.1	25.8→51.8
VGGT-1B	VFM	3	12.8	12.2→34.1	0.1	12.2→39.4	12.2→41.2
SD3.5-L, raw	DiT	7	7.9	15.4→32.6	0.1	15.4→33.4	15.4→34.8
ERNIE-Image, raw	DiT	13	13.1	37.6→54.1	0.7	37.6→59.5	37.6→59.5
SAM-B	VFM	12	3.1	26.4→40.4	0.0	26.4→40.4	26.4→41.4
CroCo-v2	VFM	8	8.2	20.2→33.4	0.1	20.2→34.1	20.2→35.5
SigLIP-B/16	VFM	8	1.2	14.5→26.5	0.0	14.5→26.5	15.7→29.7
SD2.1 (t=0)	VFM	27	0.9	12.0→23.3	0.3	12.0→24.3	13.3→26.8
Krea2, raw	DiT	13	14.6	30.2→40.0	0.9†	30.2→43.0	30.2→43.7
CleanDIFT	VFM	36	1.1	22.4→31.6	0.5	22.4→33.9	23.0→35.0
HiDream, raw	DiT	5	21.2	10.6→19.6	0.0	10.6→19.6	11.2→20.8
Qwen-Image, raw	DiT	17	8.2	22.4→29.8	0.9	22.4→33.7	22.4→35.2
SD2.1 t=250	VFM	46	0.9	5.8→12.3	0.3†	5.8→12.7	7.7→15.5
MAE-B/16	VFM	38	0.5	48.3→54.8	0.9	48.3→60.8	48.3→61.2
SD3.5-L, norm	DiT	28	0.8	24.3→30.4	0.9†	24.3→33.4	24.9→34.8
ERNIE-Image, norm	DiT	48	4.0	47.6→53.6	0.9	47.6→59.2	47.6→59.2
GLM-Image, norm	DiT	54	0.7	13.3→19.0	0.5	13.3→21.3	15.3→24.2
PE-Spatial-B	VFM	37	0.5	53.5→59.1	0.9†	53.5→61.0	53.5→61.0
Z-Image, raw	DiT	1	51.4	36.8→42.0	0.9	36.8→46.9	36.8→47.4
GLM-Image, raw	DiT	15	6.6	15.1→20.1	0.1	15.1→21.6	16.8→24.3
FLUX.1, norm	DiT	67	0.7	41.2→46.2	0.9†	41.2→50.8	41.2→51.3
V-JEPA2.1-B	VFM	33	2.8	45.5→50.4	0.9	45.5→54.5	45.5→54.8
DUNE-B/14$_{448}$	VFM	49	1.4	49.3→52.5	0.5	49.3→55.4	49.3→55.4
DUNE-B/14	VFM	49	1.1	57.8→60.4	0.5†	57.8→62.7	57.8→62.7
PixArt-α, raw	DiT	12	3.6	6.0→8.0	0.3†	6.0→8.1	7.6→10.7
FLUX2-klein, raw	DiT	94	3.0	66.9→69.0	0.9	66.9→69.0	66.9→69.0
I-JEPA-H/16	VFM	104	1.8	61.1→62.9	0.9	61.1→67.6	61.7→68.0
PixArt-α, norm	DiT	25	1.1	6.1→7.5	0.7	6.1→7.6	8.1→10.4
MiDaS-L/16	VFM	299	0.1	60.8→62.0	0.9†	60.8→63.3	60.8→63.3
Qwen-Image, norm	DiT	72	0.5	25.4→26.2	1.0	25.4→33.4	26.6→34.8
FLUX2-klein, norm	DiT	166	0.5	66.5→67.1	1.0	66.5→67.9	66.5→68.0
HiDream, norm	DiT	41	3.2	15.7→16.0	1.0	15.7→19.5	16.3→20.8
DeiT3-B/16	VFM	15	5.7	84.2→84.5	0.7	84.2→84.5	84.2→84.5
CLIP-B/16$_{\text{LAION}}$	VFM	100	0.5	76.9→77.2	0.9	76.9→77.2	76.9→77.2
Krea2, norm	DiT	93	3.9	38.4→38.5	1.0	38.4→43.1	38.4→43.9
DINO-B/16	VFM	207	0.5	76.1→76.1	0.9†	76.1→76.2	76.1→76.2
DINOv3-B/16	VFM	358	0.5	83.0→83.0	0.9	83.0→83.0	83.0→83.0
iBOT-B/16	VFM	272	0.5	77.2→77.2	0.9	77.2→77.2	77.2→77.2
DINOv2-L/14	VFM	566	0.3	83.5→83.5	1.0	83.5→83.5	83.5→83.5
DINOv2-B/14	VFM	449	0.3	82.2→82.2	0.9	82.2→82.2	82.2→82.2
DINOv3-L/16	VFM	460	0.5	85.3→85.3	0.9	85.3→85.4	85.3→85.4
Z-Image, norm	DiT	101	1.1	40.9→40.9	1.0	40.9→46.4	40.9→47.3
ConvNeXt-B	VFM	119	2.0	83.0→82.9	0.0	83.0→83.4	83.0→83.4
CLIP-B/16	VFM	100	0.4	73.9→73.7	0.3†	73.9→74.5	73.9→74.5
C-RADIOv3-B	VFM	93	0.7	77.8→77.6	0.9†	77.8→77.8	77.8→77.8
MetaCLIP2-B	VFM	92	0.3	77.3→76.6	0.3	77.3→77.8	77.3→77.8

Table 11: k-NN classification on ImageNet, all 46 readouts, sorted by gain at k=20 with the transferred α. d_{eff} and CV here are those of the representation k-NN compares (the CLS token, or the mean of the dense tokens where a model has no CLS) and not of the patch features used everywhere else in the paper. The layer is the one chosen for correspondence, not searched for classification (Section 5). α is transferred from SPair-71k validation (†: nearest grid value). The last two columns choose α on the test split, and then k as well, as Dünkel et al. (2026) do for k. The mean gain is then +7.2 and +7.7 against +5.2, and the three readouts that still fall (C-RADIOv3-B, DINOv2-B, DINOv2-L) do so by 0.05 or less.