ROTATION SIGN IN FROZEN VISION ENCODERS: WHAT TOKEN READOUTS REVEAL

Anonymous authors
Paper under double-blind review

ABSTRACT

A global image embedding can obscure information that remains accessible in the encoder's tokens. We study this distinction through rotation-sign prediction: identifying whether an image was rotated clockwise or counterclockwise. Across layers of frozen vision encoders, we compare readouts that retain token position with readouts that aggregate over it. In CLIP-B/32, a linear readout of final-layer token positions reaches 92.4% accuracy, whereas global averaging and the tested nonlinear permutation-invariant readouts remain near the 50% chance level. Training on shuffled token positions reduces accuracy to 51.2%, supporting a role for spatial organization in this contrast. DINOv2-B behaves differently: both pooled features and nonlinear invariant readouts retain substantial rotation-sign information. Additional experiments show that padding can introduce content-free cues and that probes trained on synthetic transformations do not yield a consistent advantage over permutation controls on a spatial-relation task. These results identify an encoder-dependent gap between token-level and global readouts. They also show why poor probe performance on an embedding should not be taken as evidence that the encoder has discarded the corresponding information. Establishing whether the gap is caused specifically by aggregation requires tighter capacity and optimization controls.

1 INTRODUCTION

Vision encoders are often evaluated through a single vector that summarizes an image. This interface is convenient for retrieval and classification, but it compresses a spatially organized token field. When a property cannot be recovered from the global vector, two explanations remain: the encoder may no longer represent it, or the readout may make it difficult to recover. Distinguishing these explanations matters when deciding whether a downstream failure calls for retraining the encoder or changing how its features are used.

We examine this distinction using a controlled task: predict the sign of an in-plane image rotation. Each base image is rendered with a positive or negative rotation, and a supervised probe predicts which transformation was applied. This task concerns synthetic image transformations, not three-dimensional object pose. Its simplicity makes it possible to compare layers and readouts while testing for cues introduced by rendering. We study frozen CLIP, SigLIP2, and DINO encoders (Radford et al., 2021; Tschannen et al., 2025; Oquab et al., 2023; Siméoni et al., 2025) without assuming that their training objectives make them rotation-invariant.

The main comparison holds the encoder layer fixed and changes the readout. A position-resolved probe receives tokens in their spatial order; global average pooling, attentive pooling, and DeepSets remove explicit dependence on token order. On CLIP-B/32, a linear position-resolved readout achieves 92.4% accuracy at the final layer, while the tested invariant readouts remain near chance. A probe trained on shuffled token positions also performs near chance. The information needed for this classification therefore remains accessible at the final layer, even when the tested global summaries fail to expose it.

This pattern is not universal. DINOv2-B retains substantial sign information in the average of its final-layer tokens, and nonlinear invariant readouts recover still more. Linear probes also recover sign from the native DINO embeddings, despite the weaker performance of the MLP probes used

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.

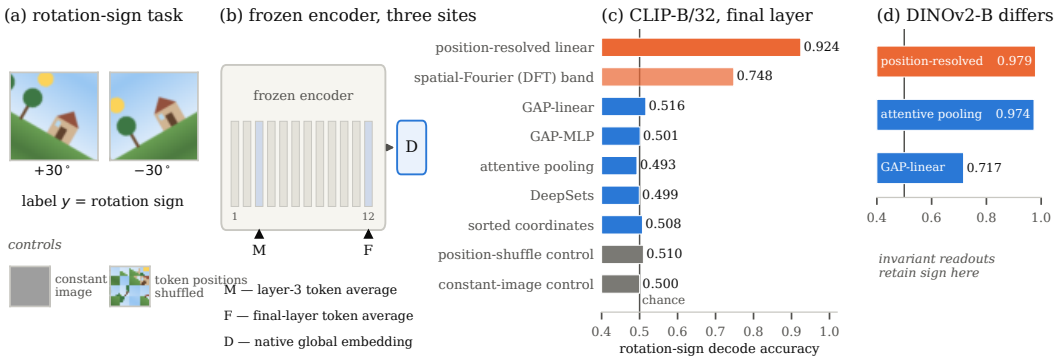

Figure 1: Study design and headline contrast. **(a)** Each base image is rendered with a positive or negative in-plane rotation under reflection padding; the label is the rotation sign (renders shown are schematic). Constant-image and token-position-shuffle controls test two possible sources of accuracy. **(b)** Frozen encoders produce token fields at each layer; we read a mid-layer average (M), a final-layer average (F), and the native global embedding (D), plus alternative readouts of the final-layer token field. **(c)** At CLIP-B/32's final layer, position-bearing readouts (orange) decode the sign while the tested permutation-invariant readouts (blue) remain near the control floor (gray). Readout parameter counts differ; Section 2 reports the relevant comparisons. **(d)** On DINOv2-B the same invariant readouts retain sign, so the pattern is encoder-dependent.

in a separate comparison. These results make the encoder and the probe part of the scientific question. They do not support a general claim that pooling removes rotation information or that a failed nonlinear probe establishes its absence.

Our contribution is an empirical characterization of this readout dependence, together with controls that delimit its interpretation. We first compare layers and readout families on CLIP-B/32 and DINOv2-B. We then examine semantic and rotation-sign prediction across five encoders, making disagreements between linear and nonlinear probes explicit. Finally, we test rendering artifacts and the transfer of transformation-derived features to spatial-relation classification. The strongest evidence concerns the accessibility of rotation sign in CLIP-B/32's final-layer tokens; broader claims about geometric mechanisms and downstream usefulness require further experiments.

2 Experimental Setup

Task and data. We construct rotation-sign examples from COCO val2017 images. For a base image x, let $T_\theta(x)$ denote its rendered in-plane rotation, and let $y \in \{-1, +1\}$ indicate the applied sign in $T_{y\theta}(x)$. The main token-readout experiment uses $m = 900$ base images, $\theta = 30°$, reflection padding, and a held-out fraction of 0.4. The experiment configuration also records a translation displacement fraction of 0.12 for the joint rotation/translation substrate. The second label, ORDER, distinguishes rotation-before-translation from translation-before-rotation; we analyze it separately because it is sensitive to boundary cues (Appendix B). Both binary tasks have chance accuracy 0.5.

Train and held-out partitions are grouped by base image. Thus transformed views of a held-out image do not enter probe training. A separate experiment balances a binary semantic label derived from COCO super-categories, leaving 864 base images. This label is unchanged by the synthetic transforms. It provides a comparison with sign prediction, although the two tasks need not have equal intrinsic difficulty. The summaries identify animal, person, and food as the positive super-category set; image-level split manifests are not included.

Encoders and feature sites. All encoder weights remain frozen; only the readouts are trained. The models are CLIP-B/32, CLIP-B/16, SigLIP2-B, DINOv2-B, and the later extension DINOv3-B. For token field $Z^{(\ell)} = (z_1^{(\ell)}, \ldots, z_T^{(\ell)})$, we consider the average $\bar{z}^{(\ell)} = T^{-1} \sum_t z_t^{(\ell)}$ at layer 3 (M) and layer 12 (F). We also evaluate the encoder's native global embedding (D). The latter is a separate readout and should not be identified with a learned projection of F without verifying the extraction implementation. Checkpoint identifiers and recorded token conventions appear in Appendix I.

Readout families and capacity. The layer sweep compares GAP-linear, GAP-MLP, position-resolved linear, single-query attentive pooling (Lee et al., 2019), DeepSets (Zaheer et al., 2017), and a spatial Fourier readout. These heads do *not* have equal parameter counts. On CLIP-B/32, GAP-linear has 769 parameters, position-resolved linear 37 633, GAP-MLP 49 346, attentive pooling 50 114, and DeepSets 57 666. The nonlinear heads share hidden width 64. The larger invariant heads are useful controls, but parameter count alone does not equalize their expressiveness or optimization.

A follow-up adds coordinate-wise sorted token values and token moments. Each is compressed by a fixed Gaussian projection before a 769-parameter linear head. These two heads match GAP-linear in trainable parameter count. Sorting separately within each feature dimension discards cross-dimensional token associations; the subsequent projection discards further information. This arm therefore does not provide an unrestricted test of the full token multiset.

Controls and uncertainty. Random-label probes provide a control-task accuracy, and selectivity is true-task accuracy minus this control (Hewitt & Liang, 2019). Because control labels are assigned to base images and evaluated on unseen base images, this is primarily a chance calibration. It does not measure training-set memorization. Constant-image controls test content-free rendering cues. Position shuffling tests dependence on token arrangement, with a separately trained shuffled-position probe in the follow-up experiment.

We report the five-seed means and 95% intervals stored in the result files. The summaries describe 2000 bootstrap draws in the main token experiments, but the supplied bundle lacks the prediction arrays and resampling implementation needed to verify dependence across transformed views, base images, and repeated seeds. We therefore treat these as reported intervals rather than independently validated uncertainty estimates. The original analyses also record Benjamini–Hochberg corrections (Benjamini & Hochberg, 1995) and decision thresholds; those records are retained in the appendix without interpreting a near-chance point estimate as proof of equivalence.

Evidence scope. The bundle contains aggregate results and the scripts that render figures and tables. It does not contain the encoder/probe execution code or the original dated design documents. Experiment identifiers in the appendix link measurements to those records; they do not by themselves verify preregistration chronology or end-to-end reproducibility. All reported experiments used development data. These constraints limit the strength of confirmatory claims.

3 Rendering controls

Synthetic transformation tasks can be solved from rendering artifacts. A rotation introduces new boundary pixels, and the padding rule can make those pixels predictive of the label. Before interpreting the layer and readout results, we therefore compare real images with constant images subjected to the same transformations. This follows the distinction between decodability and dependence on visual content emphasized by Liao et al. (2026).

On CLIP-B/32 at layer 3, both sign and transformation order are perfectly decoded from the constant-image substrate under zero padding. With reflection padding, both constant-image accuracies fall to 0.500. On real images, reflection-padded sign accuracy remains 0.967, whereas order accuracy falls from 0.925 to 0.583. The default rendering pipeline therefore supplies a strong content-free shortcut, especially for the order task.

Reflection padding removes the shortcut detected by this blank-image control. It does not show that all remaining signal is semantic or eliminate interpolation, cropping, and content-dependent boundary cues. Those alternatives require additional rendering controls. We use reflection padding for the principal token analyses and interpret the blank control narrowly: the sign result requires nonconstant image content under this renderer.

A follow-up order experiment compares real, scrambled-content, and transformation-permuted inputs. Real-image accuracy is 0.608 with reported interval $[0.558, 0.656]$, and the evidence does not resolve the source of the residual. We therefore focus the representation analysis on sign; the order results and rendering grid are retained in Appendix B.

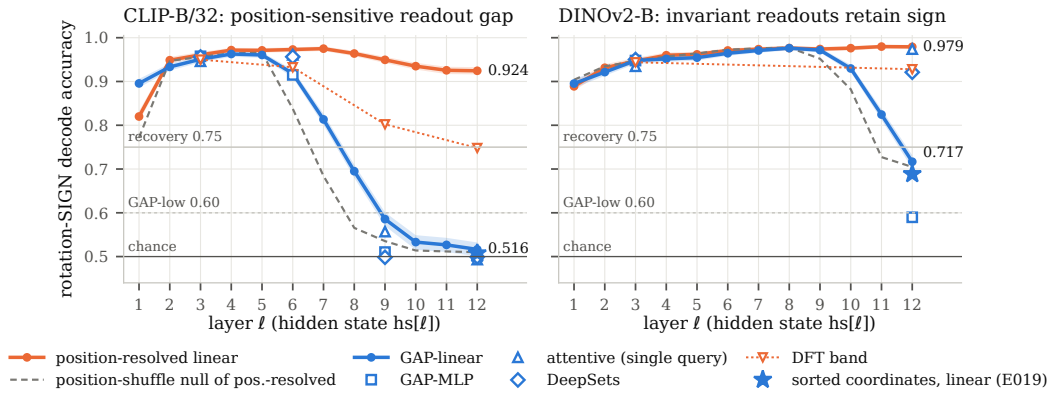

Figure 2: Rotation-sign accuracy across layers. CLIP-B/32 shows a large final-layer gap between the position-resolved probe and the tested invariant readouts. DINOv2-B retains sign information under global averaging, and nonlinear invariant readouts perform well. Bands reproduce the stored intervals; their resampling assumptions are discussed in Section 2. Horizontal lines show chance and the original descriptive thresholds.

4 Rotation sign across layers and readouts

The central experiment compares readouts of the same token field across depth. Linear heads are evaluated at each layer; nonlinear and Fourier heads are evaluated at selected layers. Figure 2 shows the trajectories, and Table 1 summarizes the final layer. Full grids, recorded parameter counts, and decision rules appear in Appendix C.

CLIP-B/32 retains an accessible positional signal. At the final layer, the position-resolved linear probe reaches 0.9244 accuracy with a reported interval of $[0.916, 0.933]$ and selectivity $+0.432$. GAP-linear reaches 0.5164, a difference of 0.408 in accuracy. The positional probe remains above 0.75 throughout the layer sweep, whereas GAP-linear falls below this level at layer 8. The result rules out complete loss of task information before the final layer: a simple supervised readout can still recover it from those tokens.

Several alternative readouts perform substantially worse at the same layer. GAP-MLP reaches 0.5014, attentive pooling 0.4931, and DeepSets 0.4992. These nonlinear heads contain more trainable parameters than the positional linear head. Their failure therefore cannot be attributed simply to having fewer parameters, although optimization and architectural restrictions remain plausible explanations. The Fourier readout reaches 0.7481; its stored adjusted p-value is 0.853 for the 0.75 recovery threshold, so it does not satisfy the original survival rule. This test concerns recovery above that threshold, not departure from chance.

Position controls support the readout contrast. The stored position-shuffle control reaches 0.510, and the constant-image control 0.500. A follow-up trains the positional head from scratch on shuffled tokens and obtains 0.512. Together these results support dependence on consistent token arrangement in the tested pipeline. They do not prove that every permutation-invariant function must fail, since finite probe families can miss information even when it is present.

Coordinate-wise sorting and token moments provide additional restricted invariant tests. Their projected linear heads reach 0.5075 and 0.5122, respectively. Neither improves on GAP-linear. Because these representations lose information before classification, this result constrains the tested summaries rather than closing the question for general set-valued readouts.

DINOv2-B retains information under invariant readouts. The corresponding DINOv2-B comparison is qualitatively different. Final-layer GAP-linear reaches 0.7167, and attentive pooling and DeepSets reach 0.9739 and 0.9211. The positional probe reaches 0.9792, but its shuffle control remains at 0.7047; a probe trained on shuffled positions reaches approximately 0.711. Consistent position is therefore not necessary for all of the detectable sign information on this encoder.

Table 1: Final-layer rotation-sign readouts. Accuracies, reported intervals, and parameter counts are read from the result files. Where counts differ by encoder, CLIP is listed first. The sorted-coordinate and moment heads use lossy fixed projections. A dash omits repeated control-head counts.

Readout	Parameters	CLIP-B/32	DINOv2-B
Permutation-invariant readouts			
GAP-linear	769	$0.516\ [0.501, 0.532]$	$0.717\ [0.701, 0.732]$
GAP-MLP	49,346	$0.501\ [0.485, 0.518]$	$0.590\ [0.574, 0.605]$
Attentive pooling	50,114	$0.493\ [0.476, 0.509]$	$0.974\ [0.969, 0.979]$
DeepSets	57,666	$0.499\ [0.484, 0.516]$	$0.921\ [0.912, 0.929]$
Sorted coordinates + projection	769	$0.508\ [0.491, 0.523]$	$0.689\ [0.672, 0.704]$
Moments + projection	769	$0.512\ [0.496, 0.529]$	$0.654\ [0.639, 0.670]$
Position-bearing readouts			
Position-resolved linear	37,633 / 49,153	$0.924\ [0.916, 0.933]$	$0.979\ [0.974, 0.984]$
Fourier band	6,913	$0.748\ [0.734, 0.762]$	$0.928\ [0.919, 0.936]$
Controls			
Position shuffle (stored)	—	0.510	0.705
Constant image	—	0.500	0.500
Trained on shuffled positions	—	$0.512\ [0.495, 0.527]$	$0.711\ [0.696, 0.726]$

The projected sorted-coordinate head reaches 0.6886, below GAP-linear, while the marginal-resampling control reaches 0.7139. These measurements are compatible with a marginal component accessible to an averaged readout and additional structure accessible to nonlinear invariant heads. They do not identify a unique carrier. In particular, preserved accuracy after resampling coordinate marginals is not a general test of whether joint token structure matters to other readouts.

The evidence supports an encoder-dependent readout gap, with a particularly clear positional contrast on CLIP-B/32. A causal attribution to aggregation alone would require comparisons that control feature dimension, training procedure, and readout capacity more tightly. The distinction between invariance to token permutations and equivariance to image rotations is also essential: the former does not establish the latter (Appendix J).

5 Native embeddings and probe dependence

The token comparison asks what can be recovered from a spatial representation. We next ask how rotation-sign prediction compares with a transform-invariant semantic task at the native global embedding. The same MLP architecture is trained separately for the two labels within each encoder. Figure 3 also shows linear sign probes, which are necessary for interpreting the nonlinear results.

Semantic accuracy remains high while MLP sign accuracy is low. On CLIP-B/32, the semantic MLP reaches 0.856, compared with 0.498 for sign. The corresponding pairs are $0.864/0.503$ on CLIP-B/16, $0.854/0.505$ on SigLIP2-B, $0.862/0.512$ on DINOv2-B, and $0.849/0.517$ on DINOv3-B. Mid-layer sign accuracy is much higher: 0.933–0.948 on the three original cross-encoder extensions. Thus the native embedding supports the semantic task well under a probe that performs poorly on sign.

Linear probes rule out a general erasure interpretation. At the same native embedding, linear sign accuracy reaches 0.639 on DINOv2-B and 0.707 on DINOv3-B, with positive reported selectivity intervals. CLIP-B/32 also has a smaller residual in the balanced semantic experiment: 0.535, with selectivity $+0.030$ and interval $[+0.007, +0.054]$. A separate sign-only run reaches 0.511. CLIP-B/16 and SigLIP2-B reach 0.529 and 0.517 in the cross-encoder experiment. These measurements establish accessible sign information in the DINO embeddings and show that the semantic-versus-sign contrast is probe-dependent.

A failed MLP does not overturn a successful linear probe. This remains true even if the MLP performs poorly in every seed: stable failure can reflect a systematic optimization or implementation

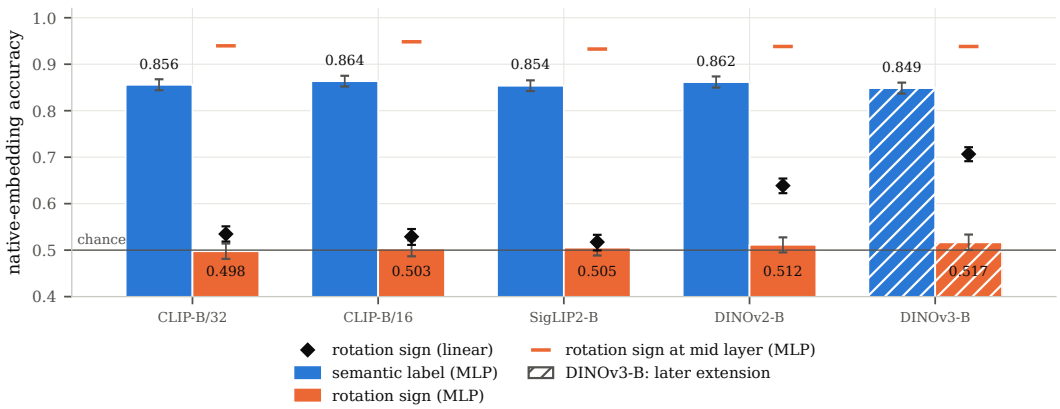

Figure 3: Native-embedding probes across encoders. Bars show the semantic and sign MLPs; diamonds show linear sign probes, and short ticks show mid-layer sign MLPs. Linear probes recover substantial sign information from both DINO embeddings. The contrast between semantic and sign MLP accuracy therefore cannot establish information loss. DINOv3-B was a later extension. Error bars reproduce the stored intervals.

problem. There is also visible seed instability at DINOv2-B's final-layer average, where GAP-MLP has mean 0.590, seed standard deviation 0.105, and individual accuracies from 0.499 to 0.708. We consequently retain the cross-encoder comparison as a description of the recorded probes, not as evidence for universal erasure or a causal effect of the pretraining objective.

The original decision rule classified four encoders as showing a semantic/sign dissociation under the specified MLP. DINOv3-B failed one clause because its sign-selectivity interval excluded zero. Appendix D preserves these outcomes. A conjunction involving a confidence interval that contains zero is not an equivalence test, and these labels do not override the positive linear results.

Dependence on rotation magnitude. A separate analysis compares the fitted sign direction at different rotation magnitudes with the direction fitted at $30°$. On CLIP-B/32 at layer 3, the reported alignment crosses zero at $49.88°$; the fitted curve places the crossing at $47.79°$. This is a descriptive regularity, not an independent forecast of that crossing: the calibration set includes $30°$, $40°$, and $50°$, so it overlaps the reported reversal. The supplied bundle also lacks the fitting implementation needed to assess the proposed functional form. We retain the measurements in Appendix E, but do not treat them as a validated geometric law.

6 Transfer to Spatial-Relation Classification

Decoding a synthetic transformation does not establish that the resulting features help another task. We test one transfer route on the controlled subsets of What's Up (Kamath et al., 2023): append two mid-layer probe logits, for sign and transformation order, to the native embedding and train a spatial-relation classifier. Subsets A and B contain 412 and 408 examples, respectively, with four relation classes. The reference arms use the native embedding alone or permuted injected features with the same added width. The injection and permutation arms have equal classifier parameter counts; the unaugmented reference does not.

The benchmark images are untransformed. The injected probes are therefore evaluated outside their synthetic training distribution, and the intervention combines sign with the unresolved order feature. It tests this two-feature transfer recipe rather than restoration of a known missing orientation variable.

No consistent advantage over the permutation reference. In the initial subset-A experiments, injection changes accuracy relative to the native-embedding baseline by -0.20 percentage points on CLIP-B/32 and -0.05 on SigLIP2-B. Relative to the permutation reference, the differences are -0.36 and -0.64 points, with reported intervals $[-0.77, -0.02]$ and $[-1.04, -0.28]$. Later runs

produce mixed results across encoder and subset. SigLIP2-B on subset B improves by $+1.23$ points over the unaugmented baseline, but its $+0.44$-point difference from the permutation reference has an interval containing zero. The full set of measurements is in Appendix G.

These results provide no consistent evidence that the particular injected features outperform their permutation controls. They do not show that rotation information is generally useless for spatial reasoning. The transformation mismatch, unresolved order feature, feature scaling, and classifier optimization remain possible explanations for weak transfer.

Oracle controls do not resolve attainable headroom. Additional arms append label-derived two-dimensional features or the one-hot relation label. They verify that informative added inputs can change predictions, but they use information unavailable at inference and therefore do not establish achievable improvement from image features. The two-dimensional oracle gains are $+3.06$ and $+4.95$ points on subset A. Their reported interval half-widths, 0.999 and 0.911 points, exceed the original 0.8-point precision target. The planned denominator criterion is consequently unresolved. Adding seeds reduced the reported intervals, but this alone does not identify the source of variance; item-level, cluster-aware uncertainty is needed before a stronger conclusion.

7 Related work

Information in frozen features. Mason et al. (2026) probe rotation across layers of large vision models using averaged patch tokens. Their results motivate a further distinction: a layer comparison with a fixed pooling rule cannot by itself determine what remains in the full token field. Our same-layer readout comparisons address that question. Other work measures geometric information in frozen representations (Shkolnikov, 2026; Gopinath et al., 2026) and recovers visual attributes through alternative token readouts (Michalkiewicz et al., 2026; Wang et al., 2026). These studies motivate separating what a backbone represents from what its default interface exposes. Our comparison uses synthetic rotation sign to examine this distinction across depth and under positional, pooled, and shuffled readouts. Its contribution is the specific controlled comparison and its encoder dependence, rather than the general observation that token features can outperform global summaries.

Invariance and readout expressiveness. Linear readouts invariant to arbitrary token permutations factor through a token sum or average. DeepSets and Set Transformers provide more expressive invariant alternatives (Zaheer et al., 2017; Lee et al., 2019). Analyses of symmetry-aware aggregation and invariant classification capacity provide a theoretical context (Talhi et al., 2026; Farrell et al., 2022), while mean-pooling studies examine related limitations in language representations (Hara et al., 2026). Our finite-probe comparisons neither prove impossibility for nonlinear invariant functions nor refute recovery guarantees whose assumptions are not met. In particular, frozen vision transformers are not assumed to realize the explicit group actions used in equivariant architectures (Cohen & Welling, 2016; Zaheer et al., 2026).

Interpreting probes and transfer. Control tasks (Hewitt & Liang, 2019) and description-length probing (Voita & Titov, 2020) emphasize that probe performance depends on the readout as well as the representation. Our random-label controls and parameter counts address parts of this issue; they are not interchangeable with a full complexity analysis. Content ablations similarly qualify what a decodable signal means (Liao et al., 2026). Prior concept-erasure attempts in this study used LEACE (Belrose et al., 2023) but failed their instrument checks, so they provide no causal support for the present interpretation. Finally, spatial-relation benchmarks expose limitations of global vision-language embeddings (Kamath et al., 2023; Yuksekgonul et al., 2023); the transfer test here evaluates only one feature-augmentation approach to that problem.

8 Discussion

The clearest result is a difference in accessibility. CLIP-B/32's final-layer tokens support accurate rotation-sign prediction when their positions are retained, while the tested global summaries perform poorly. DINOv2-B shows that this contrast depends on the encoder: invariant readouts retain

substantial information there. Poor performance on a global embedding is therefore insufficient to locate information loss within a backbone, and poor performance by one probe is insufficient to establish loss at all.

Several limitations constrain the interpretation. The central comparison uses a single synthetic transformation family over COCO images, with unequal head capacities and incomplete optimization controls. Reflection padding and constant-image controls remove one demonstrable shortcut but do not exclude every rendering cue. The cross-encoder MLP results conflict with successful linear probes, and the reported intervals require verification using base-image and seed dependence. The supplied release regenerates displays from summaries but does not yet support rerunning the experiments. No untouched test-set confirmation is available.

The most informative next experiment is a controlled readout comparison that preserves the same tokens, uses validated optimization, and measures accuracy across common capacity budgets. A second encoder should be evaluated with the same protocol regardless of its outcome. Rendering changes and a fresh dataset would test whether the positional contrast reflects a broader representation property. For transfer, an orientation-sensitive task with aligned train and test transformations would be more diagnostic than adding synthetic-task logits to untransformed images. These tests would determine whether the observed readout gap supports a general mechanism or remains a useful case study of one encoder.

AI USE STATEMENT

Generative AI tools assisted hypothesis development, experimental design, implementation, preliminary result interpretation, literature search, figure and table preparation, and manuscript drafting and revision. The rotation data were produced by deterministic image transformations rather than generative image synthesis. AI assistance was also used to revise the mathematical exposition and to audit the supplied result summaries during manuscript preparation. The authors are responsible for verifying the implementation, references, analyses, and final claims. The scope of assistance and the extent of human verification must be confirmed against the research record before submission.

ETHICS STATEMENT

This study uses public image datasets and released vision encoders. It introduces no new human-subject data collection. Public images may contain people or other sensitive content; their public availability does not remove the need to respect dataset licenses and privacy restrictions. The experiments concern synthetic in-plane transformations and should not be interpreted as measurements of real-world object pose or broad spatial reasoning ability.

REPRODUCIBILITY STATEMENT

The accompanying source contains aggregate result files and scripts that regenerate the reported tables and figures. Appendix I lists the recorded checkpoints, configurations, and provenance fields, and identifies missing execution code, split manifests, and prediction-level outputs. The current release supports checking the presentation against those summaries; it does not yet support independent execution of the experiments or verification of the bootstrap and preregistration records.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. URL https://arxiv.org/abs/1602.07576. arXiv:1602.07576.

Matthew Farrell, Blake Bordelon, Shubhendu Trivedi, and Cengiz Pehlevan. Capacity of group-invariant linear readouts from equivariant representations: How many objects can be linearly classified under all possible views? In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2110.07472. arXiv:2110.07472.

Anju Gopinath, Nikhil Krishnaswamy, and Bruce Draper. Why MLLMs struggle to determine object orientations, 2026. URL https://arxiv.org/abs/2604.13321.

Tomomasa Hara, Hiroto Kurita, Masaaki Imaizumi, Kentaro Inui, and Sho Yokoi. Why mean pooling works: Quantifying second-order collapse in text embeddings. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47180–47201, 2026. doi: 10.18653/v1/2026.acl-long.2183. URL https://aclanthology.org/2026.acl-long.2183/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. 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)*, 2019. URL https://arxiv.org/abs/1909.03368. arXiv:1909.03368.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://arxiv.org/abs/2310.19785. arXiv:2310.19785.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1810.00825. arXiv:1810.00825.

Chih-Ting Liao, Fei Shen, Xin Cao, and Tat-Seng Chua. Decodable is not grounded: A vision-ablation arbiter for VLM spatial reasoning, 2026. URL https://arxiv.org/abs/2606.31257.

Sebastian Ray Mason, Anders Gjølbye, Phillip Chavarria Højbjerg, Lenka Tětková, and Lars Kai Hansen. Large vision models can solve mental rotation problems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. URL https://arxiv.org/abs/2509.15271. arXiv:2509.15271.

Mateusz Michalkiewicz, Mahsa Baktashmotlagh, and Guha Balakrishnan. Foveated probes recover localized binding information in vision foundation models, 2026. URL https://arxiv.org/abs/2608.00726.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

Yakov Pyotr Shkolnikov. Do foundation models know geometry? probing frozen features for continuous physical measurement, 2026. URL https://arxiv.org/abs/2603.06459.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Mouad Talhi, Arne Wolf, and Anthea Monod. Breaking symmetry bottlenecks in GNN readouts, 2026. URL https://arxiv.org/abs/2602.05950.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Zhongyao Wang, Wanli Ouyang, Taoyong Cui, and Pheng Ann Heng. Support operation factorization: Compositional readout of frozen vision encoders under controlled interventions, 2026. URL https://arxiv.org/abs/2608.06174.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.01936. arXiv:2210.01936.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. URL https://arxiv.org/abs/1703.06114. arXiv:1703.06114.

Sheir A. Zaheer, Alexander C. Holston, and Chan Y. Park. REViT: Roto-reflection equivariant convolutional vision transformer. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2606.25318. arXiv:2606.25318.

A Evidence and analysis protocol

Scope of the evidence. The supplied package contains result summaries and scripts that regenerate tables and figures. It does not contain the experiment executables, per-example predictions, split manifests, or the design documents to which the summaries refer. We therefore distinguish recorded outcomes from independently reproduced findings. The reported point estimates can be checked against the summaries; the original experiments and their uncertainty estimates cannot yet be reproduced from this package alone.

Design records and hashes. Several summaries include script hashes, code revision identifiers, and substrate fingerprints. These identify the intended provenance but cannot verify it without the corresponding files. A hash establishes byte identity when both objects are available; it does not establish that a design preceded observation of the data. The original manuscript describes frozen design blocks and an access ledger, neither of which is included here. We retain experiment identifiers for traceability and describe their recorded decision rules without treating the package as independent evidence of preregistration. In particular, historical experiment counts and final-test access counts cannot be checked from the supplied files.

Recorded decisions. The summaries distinguish recovery, low accuracy, mixed results, insufficient precision, and failed validity checks. These distinctions are useful: a wide interval containing zero does not establish equivalence, and a failed gate does not establish a negative scientific result. We retain the measured values and reported flags, including conflicting flags, rather than silently replacing them with a more favorable interpretation. For example, the CLIP-B/32 E017b summaries record both a precision failure and an anomaly flag, while the SigLIP2-B denominator summaries record insufficient precision. The interpretation in Section 6 remains limited to the reported intervention.

Capacity controls. The controls differ across experiments. E014b/E016/E018 use equal input widths and equal MLP parameter counts for the semantic and transformation labels within a site. The two-feature downstream arms also have equal head counts. In E015, however, parameter counts differ across readout families: on CLIP-B/32, GAP-linear has 769 parameters, position-resolved linear 37 633, GAP-MLP 49 346, attentive pooling 50 114, DeepSets 57 666, and the DFT readout 6 913. The recorded assertion checks common hidden width and equality within a readout across layers; it does not establish equal capacity across these families. Accordingly, the positional advantage is an observed contrast between specified probes, not an isolated estimate of the effect of position at equal capacity. Larger unsuccessful nonlinear probes are informative, but their failure does not rule out optimization or architectural limitations.

Random-label controls. Selectivity is true-label accuracy minus random-label control accuracy, following the control-task approach of Hewitt & Liang (2019). Under the reported base-image grouped splits, held-out base identities and their assigned control labels are unseen during training. These controls provide a chance reference and can expose some forms of leakage; they do not reproduce a control task in which test identities are available for memorization. Neither a random-label score near chance nor a grouped split alone rules out all probe overfitting. We use selectivity with this narrower interpretation.

Uncertainty and repeated observations. All intervals in this paper are those reported in the supplied artifacts. Their coverage has not been independently verified because per-example predictions and the bootstrap implementation are absent. Rotations of the same base image, repeated predictions across seeds, and overlapping cross-validation partitions are dependent observations. A valid reanalysis must define the estimand and preserve these dependencies, resampling base images as clusters and accounting for probe-seed and fold structure where relevant. For paired contrasts, the compared predictions must remain paired during resampling.

There is a concrete reason to require this check. DINOv2-B's E016 forced-GAP MLP has five seed accuracies spanning 0.506–0.725 (SD 0.112), yet the artifact reports a 95% interval of approximately $[0.575, 0.608]$ around their mean. Moreover, the E016 linear probe uses a fixed data split and has identical accuracy across its five seed entries. Those repetitions do not supply five independent held-out datasets. The summaries alone cannot establish whether their bootstrap properly accounts for this structure; significance and precision claims remain conditional on that verification.

Multiplicity and held-out evaluation. E015 records Benjamini–Hochberg adjustments (Benjamini & Hochberg, 1995) over its SIGN cells separately for each encoder: 40 tests for CLIP-B/32 and 32 for DINOv2-B. Its recorded recovery rule requires accuracy at least 0.75, lower interval bound above 0.60, and adjusted $p < 0.05$. These are experiment-specific rules, not evidence for information-theoretic absence when a probe fails them. The adjustment does not account for subsequent selection across experiments. The supplied results are described as analyses of development data; no independent final-test confirmation is included.

Cross-condition references and unsuccessful instruments. The E015 DINOv2-B leg, E016 DINOv2-B and CLIP-B/16 legs, and E018 report using references from different conditions for a validity check. This leaves their measured accuracy unchanged but weakens the check's interpretation as reproduction under identical conditions. Earlier concept-erasure and subspace experiments (E003/E004) failed their recorded gates, and the first CLIP-B/32 E017 leg failed its probe-validity check. These files are retained for transparency and do not establish causal removal or restoration of a transformation representation.

B TRANSFORMATION ORDER AND PADDING

The auxiliary ORDER task predicts whether rotation precedes translation. Unlike rotation sign, its interpretation remains unresolved after the rendering controls. All numbers in this appendix are descriptive values from the supplied result summaries.

Rendering dependence. With zero padding, the mid-layer CLIP-B/32 probe reaches 0.925 on real images and 1.000 on constant images. Reflection padding reduces these accuracies to 0.583 and

0.500, while real-image sign accuracy remains 0.967. Thus the original order result is substantially affected by a content-free rendering cue.

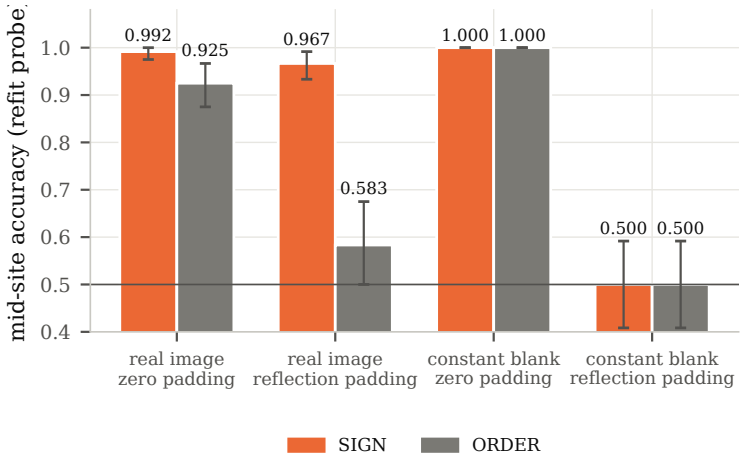

Figure 4: Content and padding controls at CLIP-B/32's mid-layer readout. Error bars reproduce the stored intervals. Reflection padding eliminates the cue detected by the constant-image control but does not exclude every content-dependent rendering artifact.

The follow-up records order accuracies of 0.608 on real images, 0.567 on scrambled content, and 0.536 on transformation-permuted pairs. The reported real-versus-scrambled difference is +0.042, with an interval including zero. This does not establish that the residual order signal is grounded in the intended image content. The exact scrambling and pair-permutation implementations are absent from the bundle.

Variation across encoders. At the native embedding, the order MLP reaches 0.855 on SigLIP2-B, 0.513 on DINOv2-B, 0.514 on CLIP-B/16, 0.513 on CLIP-B/32, and 0.924 on DINOv3-B. These differences describe probe performance; without stronger rendering controls they cannot be assigned a common geometric mechanism.

The final-layer follow-up records GAP-linear, positional, sorted-coordinate, and moment accuracies of 0.576, 0.751, 0.551, and 0.516 on CLIP-B/32; the corresponding DINOv2-B values are 0.622, 0.763, 0.604, and 0.593. These measurements remain secondary. Because the transfer experiment appends both sign and order logits, its outcome should not be described as a sign-only intervention.

An earlier order-specific gate compared reflection-padded data against a zero-padding reference and was marked invalid. Its record is retained, but it provides no evidence about the order representation.

C LAYER-BY-READOUT MEASUREMENTS

The token experiment (E015) uses 900 base images, reflection padding, 30° rotations, a 0.4 held-out fraction, and five seeds. The summaries report 720 held-out rendered examples per seed. Linear heads are measured at every layer; nonlinear and Fourier heads are measured at layers $\{3, 6, 9, 12\}$ for CLIP-B/32 and $\{3, 12\}$ for DINOv2-B. The tables retain the original threshold annotations for traceability. Their uncertainty and adjusted p-values have not been recomputed from predictions.

Capacity and preprocessing. The recorded final-layer head counts are 769 for GAP-linear, 49 346 for GAP-MLP, 50 114 for attentive pooling, 57 666 for DeepSets, and 6 913 for the Fourier head. Position-resolved linear counts differ by encoder: 37 633 for CLIP-B/32 and 49 153 for DINOv2-B. These are unequal-capacity comparisons. Shared hidden width and matching a head to its own control do not establish equality across families.

The CLIP count equals $49 \cdot 768 + 1$, consistent with a 7×7 token grid. The DINO count equals $64 \cdot 768 + 1$, whereas a full 16×16 grid would require 196 609 parameters for an unconstrained

Table 2: CLIP-B/32 rotation-sign accuracy. Bold entries meet all stored survival criteria: accuracy ≥ 0.75, interval lower bound > 0.60, and BH-adjusted $p < 0.05$. A dagger marks accuracy ≥ 0.75 that fails another criterion; a dash denotes an unmeasured cell.

Layer	GAP-lin.	GAP-MLP	attentive	DeepSets	pos.-res. lin.	shuffle null	DFT band	DFT shuffle
hs[1]	**0.896**	—	—	—	**0.820**	0.772	—	—
hs[2]	**0.933**	—	—	—	**0.949**	0.947	—	—
hs[3]	**0.951**	**0.958**	**0.946**	**0.956**	**0.961**	0.956	**0.950**	0.949
hs[4]	**0.963**	—	—	—	**0.972**	0.964	—	—
hs[5]	**0.961**	—	—	—	**0.971**	0.966	—	—
hs[6]	**0.918**	**0.915**	**0.961**	**0.956**	**0.973**	0.838	**0.932**	0.754
hs[7]	**0.813**	—	—	—	**0.975**	0.684	—	—
hs[8]	0.695	—	—	—	**0.964**	0.566	—	—
hs[9]	0.586	0.510	0.557	0.498	**0.949**	0.535	**0.802**	0.526
hs[10]	0.533	—	—	—	**0.935**	0.514	—	—
hs[11]	0.527	—	—	—	**0.926**	0.512	—	—
hs[12]	0.516	0.501	0.493	0.499	**0.924**	0.510	0.748	0.510

Table 3: DINOv2-B rotation-sign accuracy, using the same annotations. Final-layer invariant readouts retain substantial information.

Layer	GAP-lin.	GAP-MLP	attentive	DeepSets	pos.-res. lin.	shuffle null	DFT band	DFT shuffle
hs[1]	**0.895**	—	—	—	**0.889**	0.904	—	—
hs[2]	**0.921**	—	—	—	**0.931**	0.935	—	—
hs[3]	**0.948**	**0.948**	**0.935**	**0.950**	**0.947**	0.947	**0.944**	0.938
hs[4]	**0.952**	—	—	—	**0.960**	0.957	—	—
hs[5]	**0.954**	—	—	—	**0.962**	0.964	—	—
hs[6]	**0.964**	—	—	—	**0.971**	0.972	—	—
hs[7]	**0.971**	—	—	—	**0.974**	0.976	—	—
hs[8]	**0.976**	—	—	—	**0.976**	0.978	—	—
hs[9]	**0.972**	—	—	—	**0.974**	0.952	—	—
hs[10]	**0.930**	—	—	—	**0.976**	0.882	—	—
hs[11]	**0.824**	—	—	—	**0.980**	0.727	—	—
hs[12]	0.717	0.590	**0.974**	**0.921**	**0.979**	0.705	**0.928**	0.634

scalar linear head. This arithmetic establishes a mismatch between the reported native grid and a naive flattened head, but it does not identify the unrecorded preprocessing. Spatial downsampling is one possible explanation, not a verified implementation detail.

Shuffle controls. The original summary does not specify whether the E015 positional shuffle is applied during training or only during evaluation. E019 records a separately trained shuffled-position head, with accuracy approximately 0.512 on CLIP-B/32 and 0.711 on DINOv2-B. These are close to the respective E015 controls (0.510 and 0.705). The comparison supports a dependence on consistent positions for the CLIP pipeline, while the DINO pipeline preserves a substantial component recoverable without consistent token slots.

Recorded decision labels. The stored E015 outcomes are INTERFACE-POSITIONAL for CLIP-B/32 and MIXED for DINOv2-B. DINO's GAP accuracy never enters the rule's ≤ 0.60 regime, and its position-shuffle control remains above that threshold. We retain these labels as analysis metadata rather than treating them as proof of an exhaustive depth-versus-pooling dichotomy. The DINO leg also uses a cross-condition gate reference (Appendix A).

D CROSS-ENCODER PROBE MEASUREMENTS

The semantic/sign comparison combines E014b, E016, and the later DINOv3-B extension E018. The recorded semantic balancing leaves 864 base images, with 518 training and 346 held-out base images. Shared settings include 30° rotations, reflection padding, five seeds, and hidden width 64 for the MLPs. These are equal architectures for the two labels within a given site, not equal parameter counts across all encoders and sites.

Table 4: Semantic and sign MLP measurements, with native-embedding linear sign probes. The MLP-conjunction column reproduces the original MLP conjunction; it does not certify information loss. The half-width column describes the stored sign interval.

Encoder	Sign @M MLP	Semantic @D MLP accuracy [95% CI]	Sign @D MLP accuracy [95% CI]	Sign @D Linear
CLIP-B/32	0.940	0.856 [0.844, 0.868]	0.498 [0.481, 0.514]	0.535
CLIP-B/16	0.948	0.864 [0.852, 0.875]	0.503 [0.487, 0.521]	0.529
SigLIP2-B	0.933	0.854 [0.842, 0.865]	0.505 [0.488, 0.522]	0.517
DINOv2-B	0.938	0.862 [0.850, 0.874]	0.512 [0.495, 0.527]	0.639
DINOv3-B	0.938	0.849 [0.837, 0.860]	0.517 [0.500, 0.534]	0.707

Encoder	Semantic @D Selectivity CI	Sign @D Selectivity CI	MLP conjunction	Cross-cond. gate	Sign @D half-width
CLIP-B/32	$[+0.339, +0.380]$	$[-0.022, +0.027]$	yes	no	—
CLIP-B/16	$[+0.344, +0.385]$	$[-0.018, +0.028]$	yes	yes	0.0171
SigLIP2-B	$[+0.333, +0.375]$	$[-0.018, +0.031]$	yes	no	0.0166
DINOv2-B	$[+0.327, +0.368]$	$[-0.025, +0.021]$	yes	yes	0.0162
DINOv3-B	$[+0.344, +0.385]$	$[+0.008, +0.056]$	no	yes	0.0166

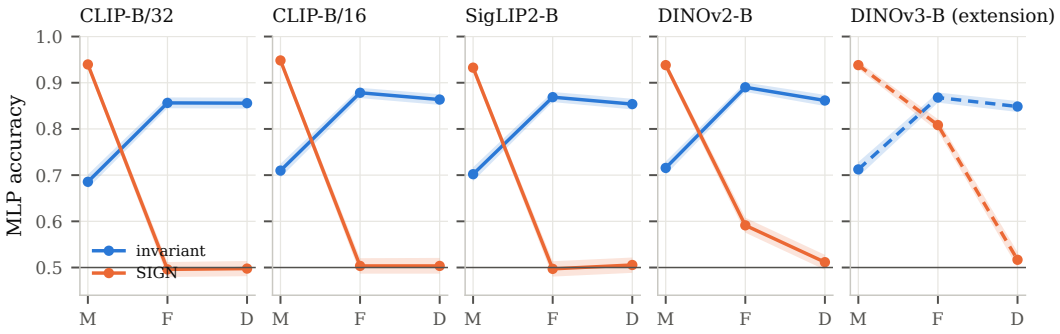

Figure 5: MLP accuracy at the mid-layer average (M), final-layer average (F), and native embedding (D). These are separate feature extractions; the diagram is not evidence that D is computed by projecting F. Bands are stored intervals.

Linear evidence takes precedence over failed decoding. Native-embedding linear sign accuracy is 0.535 on CLIP-B/32, 0.529 on CLIP-B/16, 0.517 on SigLIP2-B, 0.639 on DINOv2-B, and 0.707 on DINOv3-B. DINOv2 and DINOv3 have positive selectivity intervals of $[+0.114, +0.162]$ and $[+0.188, +0.233]$. CLIP-B/32's smaller residual also has positive reported selectivity on the balanced substrate. These results contradict a universal absence-of-information interpretation, regardless of the original MLP outcome flags.

Seed variation and systematic probe failure. For DINOv2-B's final-layer average, E015 records GAP-MLP seed accuracies of $\{0.499, 0.700, 0.528, 0.514, 0.708\}$, with standard deviation 0.105. E016 records a similar pattern and standard deviation 0.112. The narrow stored item-level intervals do not capture this variation by themselves. Native-embedding MLP seed standard deviations are smaller (0.007–0.019), but low dispersion cannot exclude a systematic training error. A diagnostic implementation should first verify that an MLP can reproduce the trained linear classifier before using MLP failure to interpret a representation.

Historical decision rules. The original conjunction required semantic MLP accuracy at least 0.75, a semantic interval lower bound above 0.60, positive semantic selectivity, sign accuracy at most 0.60, and a sign-selectivity interval containing zero. Four encoders satisfy this recorded conjunction. DINOv3-B fails its last clause: the reported sign-selectivity interval is $[+0.0084, +0.0561]$. This is a probe-specific outcome, not a fourth or fifth demonstration of erasure. Nor does the DINOv2 MLP

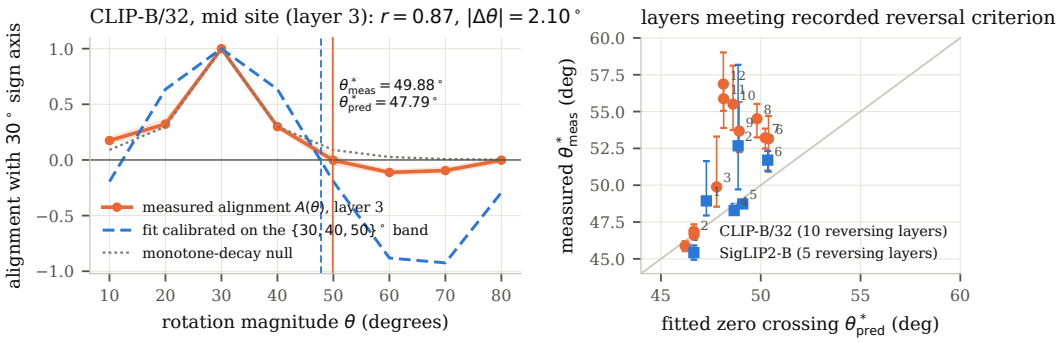

Figure 6: Rotation-magnitude analysis from E013. **Left**: CLIP-B/32 layer-3 alignment, the stored curve calibrated on $\{30, 40, 50\}°$, and the stored positive-decay comparison. The highlighted crossing lies within the calibration band. **Right**: estimated and fitted crossings for layers meeting the recorded reversal and precision criteria; labels identify layers, and bars show reported bootstrap CIs. The full table also includes layers that do not meet these criteria.

result falsify an objective-level retention hypothesis in the presence of a successful linear probe. Architectures, data, and training objectives are not independently varied here.

DINOv3's blank semantic accuracies (0.539–0.564) are also higher than the approximately 0.500 controls in most earlier runs. They remain below the stored gate but warrant verification against the balanced label and split implementation. The original flags and all per-site values remain available in the unchanged JSON files.

E PROBE DIRECTIONS ACROSS ROTATION MAGNITUDES

This supplementary analysis asks whether a rotation-sign probe fitted at one angle has the same direction as probes fitted at other angles. E013 uses pure rotations $\pm\theta$, reflection padding, $m = 900$ base images, a held-out fraction of 0.4, and five seeds. The angle grid is $\{10, 20, \ldots, 80\}°$ and the reference angle is $\theta_{\text{ref}} = 30°$. For each layer ℓ, the recorded alignment is

$$A_\ell(\theta) = \frac{w_\ell(\theta)^\top w_\ell(30°)}{\|w_\ell(\theta)\| \, \|w_\ell(30°)\|},$$

where $w_\ell(\theta)$ is a logistic-regression coefficient vector fitted to rotation sign at magnitude θ. Negative alignment means that the two fitted directions form an obtuse angle; it does not imply that the vectors are negatives of one another or that their downstream effects reverse.

Recorded fits and zero crossings. The experiment summary records a per-layer constant k_ℓ, fitted curves, and their predicted zero crossings. It names $\{30, 40, 50\}°$ as the calibration band. The exact fitting objective and generating implementation are not included in the source package, so the renderer reproduces the stored curves rather than refitting them. The recorded targets were crossing error $\leq 15°$, curve correlation $r \geq 0.8$, crossing CI half-width $\leq 12°$, and positive-control accuracy ≥ 0.75.

Table 5 retains every layer. The artifact marks a layer as showing a "reversal" when the upper confidence limit of alignment is negative at at least one sampled angle. This criterion is met at 10 CLIP-B/32 layers and 5 SigLIP2-B layers. At CLIP-B/32 layer 3, the estimated zero crossing is $49.88°$ $[48.54, 53.30]$, compared with the fitted curve's crossing at $47.79°$, an error of $2.10°$ and a whole-grid correlation of 0.87. The recorded random-axis control is 0.012 $[-0.008, 0.039]$, the shuffled-token floor is 0.512, and the blank control is 0.500; the two encoder positive controls are 0.951 and 0.958. The original outcome map records K_1 for these results. That historical label is retained without treating it as independent validation of a predictive law.

Limits of the prediction claim. The highlighted $49.88°$ crossing falls within the 30–50° calibration band, so agreement with $47.79°$ is not a prediction of a withheld crossing. Correlation also does

Table 5: E013 per-layer summaries. "Reversal" denotes the recorded negative-alignment CI criterion; "precision" denotes the crossing half-width criterion. A dash indicates no finite crossing of the mean curve. Finite crossings in layers without a reversal flag should not be interpreted as statistically established negative alignment.

Layer	reversal	θ^*_{meas} [CI]	θ^*_{pred}	$\lvert\Delta\theta\rvert$	r	precision
CLIP-B/32: reversing layers 1, 2, 3, 6, 7, 8, 9, 10, 11, 12; gate 0.951						
1	yes	45.88 [45.53, 46.20]	46.22	0.34	0.822	yes
2	yes	46.64 [46.32, 47.13]	46.66	0.02	0.803	yes
3	yes	49.88 [48.54, 53.30]	47.79	2.10	0.870	yes
4	no	—	49.20	—	0.877	no
5	no	—	51.12	—	0.909	no
6	yes	53.16 [50.91, 54.70]	50.39	2.76	0.954	yes
7	yes	53.22 [52.49, 53.84]	50.23	2.99	0.958	yes
8	yes	54.51 [53.25, 55.52]	49.81	4.70	0.944	yes
9	yes	53.67 [52.25, 55.64]	48.92	4.75	0.909	yes
10	yes	55.51 [53.75, 58.12]	48.61	6.89	0.841	yes
11	yes	55.87 [53.89, 56.92]	48.14	7.73	0.804	yes
12	yes	56.87 [55.05, 59.02]	48.13	8.74	0.809	yes
SigLIP2-B: reversing layers 1, 2, 4, 5, 6; gate 0.958						
1	yes	48.94 [47.95, 51.63]	47.27	1.66	0.834	yes
2	yes	52.68 [49.72, 58.17]	48.87	3.81	0.912	yes
3	no	—	49.95	—	0.913	no
4	yes	48.28 [48.01, 48.73]	48.67	0.39	0.918	yes
5	yes	48.73 [48.55, 48.92]	49.10	0.37	0.929	yes
6	yes	51.70 [50.99, 52.32]	50.35	1.35	0.952	yes
7	no	66.67 [64.51, 69.18]	52.15	14.52	0.951	yes
8	no	—	50.51	—	0.913	no
9	no	—	50.61	—	0.864	no
10	no	69.84 [60.59, 69.94]	49.20	20.64	0.862	no
11	no	69.27 [63.08, 69.72]	49.42	19.86	0.858	yes
12	no	—	49.73	—	0.870	no

not measure agreement in amplitude. At 60° and 70°, for example, the stored CLIP-B/32 layer-3 curve predicts approximately -0.881 and -0.925, while the observed alignments are -0.112 and -0.095. The artifact's residual fields cannot be independently reconstructed without the fitting code and their precise definitions. A positive-decay model cannot cross zero, but its failure to represent a negative value does not establish superior prediction error for the alternative curve. SigLIP2-B also has finite mean-curve crossings around 67–70° at layers that fail the recorded reversal criterion; these remain in Table 5 rather than in the selected-layer plot.

The supported observation is that fitted probe directions vary with rotation magnitude, with negative alignment at some angles and layers. Establishing a predictive model requires its formula and fitting code, a calibration interval strictly separated from the evaluation angles, and comparisons of held-out prediction error. Cross-angle classification is also needed before interpreting coefficient alignment as transfer of rotation-sign decoding. E013 does not measure deployed embeddings, the ORDER label, or the effect of steering a downstream representation.

F SORTED-COORDINATE AND MOMENT READOUTS

E019 compares six final-layer arms: GAP-linear, a positional linear head, a separately trained positional head on shuffled tokens, sorted-coordinate features, token moments, and a coordinate-marginal resampling control. The sorted and moment features pass through a fixed seeded Gaussian projection before a linear head. GAP-linear and these two heads each have 769 trainable parameters.

What sorting preserves. Sorting each coordinate across tokens preserves its empirical marginal values and removes explicit token order. It does not preserve the associations among coordinates within a token. For example, the token sets $\{(0,0),(1,1)\}$ and $\{(0,1),(1,0)\}$ have identical sorted coordinates but different joint structure. The Gaussian projection is a further compression. Conse-

Table 6: Final-layer follow-up measurements, with five seeds and 720 held-out rendered examples per seed. The original decision thresholds apply to the sorted-coordinate arm; intervals are reproduced from the artifacts.

Arm	SIGN acc. [CI]	Selectivity [CI]	shuffle null	params
CLIP-B/32: frozen outcome NEGATIVE branch did not fire; $d - a = -0.0089$; $e - d = +0.0047$; hw(d) $= 0.0161$				
(a) GAP-linear	0.516 [0.501, 0.532]	+0.011 [−0.012, +0.035] (incl. 0)	—	769
(b) position-resolved linear	0.924 [0.916, 0.933]	+0.432 [+0.414, +0.451] (excl. 0)	0.510	37633
(c) position-shuffled input to (b)	0.512 [0.495, 0.527]	+0.018 [−0.006, +0.041] (incl. 0)	—	37633
(d) sorted coords. + projection, linear	0.508 [0.491, 0.523]	+0.013 [−0.010, +0.036] (incl. 0)	—	769
(e) token moments, linear	0.512 [0.496, 0.529]	+0.002 [−0.020, +0.024] (incl. 0)	—	769
(f) token-marginal-resample null	0.508 [0.492, 0.524]	+0.005 [−0.019, +0.027] (incl. 0)	—	37633
DINOv2-B: frozen outcome K_3, intermediate; $d - a = -0.0281$; $e - d = -0.0342$; hw(d) $= 0.0160$				
(a) GAP-linear	0.717 [0.701, 0.732]	+0.220 [+0.198, +0.243] (excl. 0)	—	769
(b) position-resolved linear	0.979 [0.974, 0.984]	+0.464 [+0.446, +0.480] (excl. 0)	0.705	49153
(c) position-shuffled input to (b)	0.711 [0.696, 0.726]	+0.206 [+0.184, +0.229] (excl. 0)	—	49153
(d) sorted coords. + projection, linear	0.689 [0.672, 0.704]	+0.182 [+0.160, +0.205] (excl. 0)	—	769
(e) token moments, linear	0.654 [0.639, 0.670]	+0.146 [+0.124, +0.169] (excl. 0)	—	769
(f) token-marginal-resample null	0.714 [0.699, 0.730]	+0.207 [+0.184, +0.230] (excl. 0)	—	49153

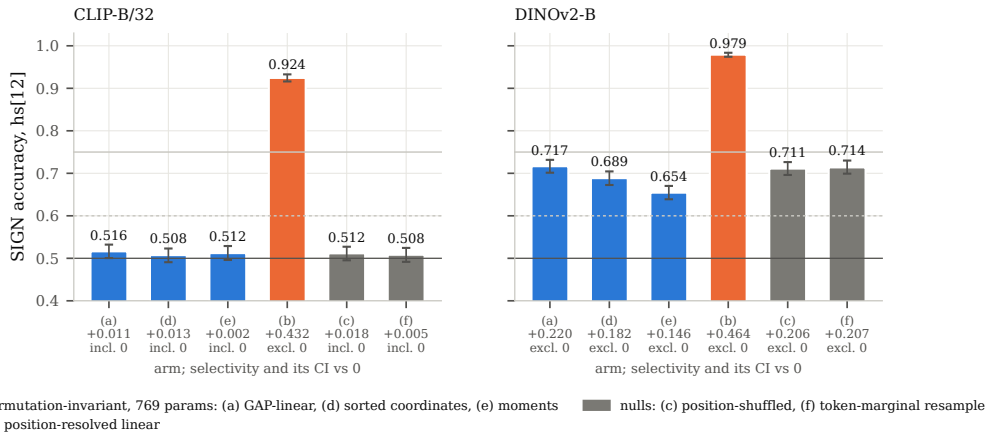

Figure 7: Readouts and shuffle/resampling controls. On CLIP-B/32 the tested invariant summaries perform near chance. DINOv2-B retains a component accessible to GAP and to the marginal-resampling pipeline. These arms do not exhaust invariant functions of the token field.

quently, this representation must not be described as an information-preserving encoding of the full token multiset.

The CLIP sorted-coordinate and moment accuracies are 0.5075 and 0.5122, compared with 0.5164 for GAP-linear. This extends the observed failure to two more restricted summaries. It does not prove that a sufficiently trained invariant set model could not recover sign.

DINOv2's sorted-coordinate accuracy is 0.6886, below GAP-linear at 0.7167; its marginal-resampling control reaches 0.7139. An operation that preserves empirical coordinate marginals exactly also preserves their means; independent resampling preserves them only in expectation. A mean-based readout can therefore remain informative under this control. The exact resampling-to-readout path must be recovered from the missing execution code before attributing additional mechanistic meaning. Attentive and DeepSets accuracies of 0.9739 and 0.9211 establish that the tested nonlinear invariant heads can exploit more sign-predictive structure on this encoder.

The stored DINO outcome is intermediate (K_3). The CLIP sorted-coordinate arm does not cross the original recovery threshold. We preserve these records without treating their branch names as a classification of all possible signal carriers.

G FEATURE INJECTION AND ORACLE CONTROLS

E005 and E006 append two mid-layer transformation logits to What's Up embeddings. A four-class linear head predicts spatial relations. The no-op arm uses the native embedding, the permutation arm uses the same added width with the recorded permutation control, and an unmatched reference appends the full mid-layer representation. Later E017/E017b experiments add constant, random-projection, task-direction, and label-derived oracle inputs. The summaries do not include the exact estimators for the task-direction arms or the full permutation implementation.

Table 7: Initial feature-injection results by encoder and subset. Gains and intervals are reproduced from the aggregate result files.

Encoder	subset	n	no-op	injected	perm. null	K_1 (pp)	K_2 (pp)	unmatched mid (pp)
CLIP-B/32 (E005)	A	412	0.8022	0.8002	0.8039	-0.20 $[-0.59, +0.11]$	-0.36 $[-0.77, -0.02]$	$+12.88$ $[+11.39, +14.52]$
CLIP-B/32 (E005)	B	408	0.7529	0.7514	0.7531	-0.14 $[-0.78, +0.46]$	-0.17 $[-0.81, +0.41]$	$+7.86$ $[+6.00, +9.75]$
SigLIP2-B (E006)	A	412	0.8035	0.8030	0.8094	-0.05 $[-0.25, +0.19]$	-0.64 $[-1.04, -0.28]$	$+15.95$ $[+14.43, +17.46]$
SigLIP2-B (E006)	B	408	0.7416	0.7588	0.7528	$+1.71$ $[+0.94, +2.48]$	$+0.60$ $[-0.14, +1.34]$	$+10.29$ $[+8.95, +11.68]$

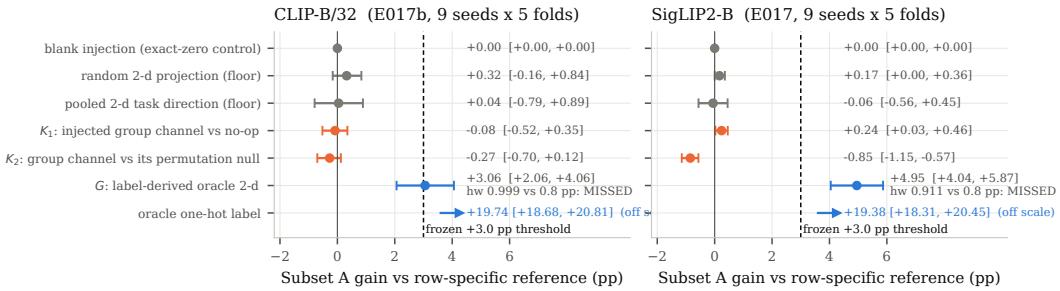

Figure 8: Subset-A injection and oracle contrasts. Each row uses the reference named in its label; the permutation contrast is not a no-op contrast. Label-derived oracle inputs are unavailable at inference. Their gains demonstrate sensitivity to added label information, not attainable visual headroom.

Capacity and distribution. CLIP-B/32 uses 512 native features, 514 inputs for the two-feature arms, and 516 for the one-hot oracle. The two-feature heads each have 2060 parameters; the other heads differ. The benchmark images are not synthetically transformed, so both injected logits are evaluated outside the distribution on which their labels were defined. In addition, order grounding remains unresolved. The experiment therefore concerns the specific two-logit feature augmentation, not a general intervention restoring rotation sign.

Mixed transfer outcomes. In the later SigLIP2-B subset-A run, injection gains $+0.24$ points over no-op, compared with $+0.17$ for random features and $+1.09$ for the permutation arm. On subset B, injection gains $+1.23$ points over no-op, with reported interval $[+0.70, +1.77]$, but its $+0.44$-point permutation contrast includes zero. CLIP-B/32 subset B has a negative permutation contrast of -0.59 points with interval $[-1.05, -0.16]$. These differences motivate reporting all references rather than selecting the most favorable comparison. The cause of the permutation arm's positive no-op offset is unknown; regularization is only one possible explanation.

Oracle precision. On subset A, two-dimensional oracle gains are $+3.06$ points for CLIP-B/32 and $+4.95$ for SigLIP2-B. Their half-widths are 0.999 and 0.911 points after the extension, exceeding the 0.8-point target. CLIP's interval $[+2.06, +4.06]$ also crosses the original $+3.0$-point threshold. These results do not satisfy the planned denominator criterion. Because the oracle uses label-derived features, satisfying the numerical criterion would still not demonstrate a realizable visual improvement.

Increasing seeds from five to nine reduced the reported half-widths, but does not by itself show that uncertainty is dominated by held-out images. A dependence-aware analysis should resample independent examples or acquisition groups and account for repeated predictions and overlapping folds. The summaries are insufficient to perform that analysis.

Table 8: Later experiments after the nine-seed, five-fold extension. All recorded contrasts are shown for both subsets. The 0.8-point half-width target applies to the two-dimensional oracle gain.

Quantity	gain vs no-op (pp) [CI]	half-width	CI vs 0	hw vs 0.8 pp target
CLIP-B/32, E017b, 9×5, What's Up subset A ($n = 412$; no-op 0.8023)				
blank injection (exact-zero control)	$+0.00$ $[+0.00, +0.00]$	0.000	incl. 0	
random 2-d projection (floor)	$+0.32$ $[-0.16, +0.84]$	0.499	incl. 0	
pooled 2-d task direction (floor)	$+0.04$ $[-0.79, +0.89]$	0.842	incl. 0	
K_1 injected group channel vs no-op	-0.08 $[-0.52, +0.35]$	0.435	incl. 0	
K_2 group channel vs permutation null	-0.27 $[-0.70, +0.12]$	0.411	incl. 0	
sign-only injection vs no-op	-0.05 $[-0.43, +0.32]$	0.378	incl. 0	
order-only injection vs no-op	-0.13 $[-0.44, +0.14]$	0.290	incl. 0	
permutation null vs no-op	$+0.19$ $[-0.04, +0.42]$	0.233	incl. 0	
G oracle 2-d task direction (ceiling)	$+3.06$ $[+2.06, +4.06]$	0.999	excl. 0	MISSED
oracle one-hot label (ceiling)	$+19.74$ $[+18.68, +20.81]$	1.062	excl. 0	
unmatched mid-layer features (reference)	$+12.73$ $[+11.59, +13.88]$	1.145	excl. 0	
CLIP-B/32, E017b, 9×5, What's Up subset B ($n = 408$; no-op 0.7480)				
blank injection (exact-zero control)	$+0.00$ $[+0.00, +0.00]$	0.000	incl. 0	
random 2-d projection (floor)	$+1.26$ $[+0.69, +1.84]$	0.577	excl. 0	
pooled 2-d task direction (floor)	-1.09 $[-1.84, -0.30]$	0.768	excl. 0	
K_1 injected group channel vs no-op	-0.28 $[-0.74, +0.14]$	0.439	incl. 0	
K_2 group channel vs permutation null	-0.59 $[-1.05, -0.16]$	0.442	excl. 0	
sign-only injection vs no-op	-0.03 $[-0.44, +0.37]$	0.401	incl. 0	
order-only injection vs no-op	-0.06 $[-0.47, +0.34]$	0.407	incl. 0	
permutation null vs no-op	$+0.31$ $[-0.01, +0.63]$	0.319	incl. 0	
G oracle 2-d task direction (ceiling)	$+2.19$ $[+1.28, +3.09]$	0.903	excl. 0	MISSED
oracle one-hot label (ceiling)	$+25.12$ $[+23.79, +26.47]$	1.336	excl. 0	
unmatched mid-layer features (reference)	$+7.96$ $[+6.65, +9.22]$	1.284	excl. 0	
SigLIP2-B, E017, 9×5, What's Up subset A ($n = 412$; no-op 0.8038)				
blank injection (exact-zero control)	$+0.00$ $[+0.00, +0.00]$	0.000	incl. 0	
random 2-d projection (floor)	$+0.17$ $[+0.00, +0.36]$	0.177	excl. 0	
pooled 2-d task direction (floor)	-0.06 $[-0.56, +0.45]$	0.507	incl. 0	
K_1 injected group channel vs no-op	$+0.24$ $[+0.03, +0.46]$	0.215	excl. 0	
K_2 group channel vs permutation null	-0.85 $[-1.15, -0.57]$	0.289	excl. 0	
sign-only injection vs no-op	$+0.03$ $[-0.08, +0.14]$	0.108	incl. 0	
order-only injection vs no-op	$+0.21$ $[+0.03, +0.41]$	0.190	excl. 0	
permutation null vs no-op	$+1.09$ $[+0.85, +1.35]$	0.251	excl. 0	
G oracle 2-d task direction (ceiling)	$+4.95$ $[+4.04, +5.87]$	0.911	excl. 0	MISSED
oracle one-hot label (ceiling)	$+19.38$ $[+18.31, +20.45]$	1.070	excl. 0	
unmatched mid-layer features (reference)	$+16.14$ $[+15.02, +17.23]$	1.107	excl. 0	
SigLIP2-B, E017, 9×5, What's Up subset B ($n = 408$; no-op 0.7436)				
blank injection (exact-zero control)	$+0.00$ $[+0.00, +0.00]$	0.000	incl. 0	
random 2-d projection (floor)	$+0.24$ $[-0.25, +0.74]$	0.495	incl. 0	
pooled 2-d task direction (floor)	-1.67 $[-2.27, -1.05]$	0.610	excl. 0	
K_1 injected group channel vs no-op	$+1.23$ $[+0.70, +1.77]$	0.538	excl. 0	
K_2 group channel vs permutation null	$+0.44$ $[-0.04, +0.88]$	0.462	incl. 0	
sign-only injection vs no-op	$+0.05$ $[-0.36, +0.43]$	0.396	incl. 0	
order-only injection vs no-op	$+0.78$ $[+0.26, +1.27]$	0.504	excl. 0	
permutation null vs no-op	$+0.79$ $[+0.47, +1.11]$	0.318	excl. 0	
G oracle 2-d task direction (ceiling)	$+5.13$ $[+4.24, +6.02]$	0.893	excl. 0	MISSED
oracle one-hot label (ceiling)	$+24.80$ $[+23.52, +26.31]$	1.396	excl. 0	
unmatched mid-layer features (reference)	$+10.11$ $[+9.02, +11.30]$	1.140	excl. 0	

Recorded exclusions and labels. The first CLIP leg of E017 used an order-validity threshold inherited from a different padding condition and was excluded. E017b is the renamed successor. Its raw scorer emits `MIXED_ANOMALY` because a zero-gain blank arm exceeds a slightly negative injected gain. This bit is preserved in the artifact. We report the numerical precision failure directly rather than rewriting the stored output or interpreting this branch as evidence of leakage.

H Experiment index and historical outcomes

Table 9 links the principal analyses to their original identifiers. The decision labels are historical metadata from the supplied summaries and notes. The original dated design blocks are not included, so this release cannot independently verify that each rule preceded the corresponding measurements.

The main text interprets the quantitative results without promoting the branch names to scientific conclusions.

Table 9: Experiment index. All listed interpretations are bounded by the provenance and uncertainty limitations in Appendices A and I.

Record	Question and recorded outcome
E005/E006	Transfer of sign/order logits; initial subset-A permutation contrasts are negative.
E007/E008a	Content and padding controls; zero padding permits content-free prediction.
E009	Residual order grounding; the recorded analysis is inconclusive.
E011	Second-order token readout; final-layer sign accuracy 0.567, below the original 0.75 recovery threshold.
E012	Probe complexity; native-embedding sign MLP accuracy 0.501 on CLIP-B/32.
E013	Magnitude dependence; recorded reversal/fitted-curve outcome K_1, qualified by calibration overlap and unspecified fit evaluation.
E014b/E016	Semantic/sign MLP comparison; four encoders satisfy the recorded conjunction, with positive linear sign results on DINOv2-B.
E015	Layer/readout sweep; CLIP-B/32 labeled INTERFACE-POSITIONAL, DINOv2-B MIXED.
E017/E017b	Additional transfer controls; planned oracle precision not met on either encoder.
E018	DINOv3-B extension; semantic/sign MLP contrast present, original conjunction fails on sign selectivity.
E019	Sorted-coordinate and moment readouts; DINOv2-B assigned intermediate outcome K_3.

Excluded instruments and runs. The supplied notes describe two E003 concept-erasure attempts that failed transfer-validity checks, E004's invalid low-rank injection instrument, E010's incompatible order/padding gate, E014's semantic gate at the wrong feature site, and the first CLIP leg of E017. Available excluded-run JSON files are retained in the release. The narrative invalid counter and enumerated run history do not fully reconcile, and not every named run has an individual JSON file. These limitations preclude a claim that the release contains a complete execution history.

No results from these excluded instruments support the main conclusions. Successor experiments have distinct identifiers in the notes; confirming their chronology requires the original run documents and source archive.

I Reproducibility and artifact inventory

What the package reproduces. The source includes 29 JSON result files, LaTeX tables, PDF figures, and scripts for checking numerical references and regenerating tables and figures. Two E014b filenames contain identical bytes and describe the same run. This package reproduces the presentation of recorded summaries. It does not currently reproduce feature extraction, probe fitting, transformations, or statistical inference from individual predictions.

Datasets and sample sizes. The summaries identify COCO val2017 as the source of base images. E014b/E016/E018 also use its instance annotations for a balanced binary label based on supercategories; the positive set is animal, person, and food. These experiments retain 864 base images, with 518 training and 346 held-out base images. Their binary rotation tasks contain two variants per image, giving $1\,036$ training and 692 held-out examples. E015/E019 use 900 base images with a held-out fraction of 0.4, giving 720 held-out rotation variants per seed. Earlier grounding and injection runs E005–E008a use 300 base images and a held-out fraction of 0.2; E009 uses 900 and 0.2. E017/E017b also use 900 and 0.2 for the synthetic probe-training substrate. These are distinct from the downstream What's Up subsets (412 and 408 images). The transformations use $\theta = 30^\circ$ and, where translation is used, displacement fraction 0.12. Reflection padding applies to the later controlled experiments; the early grounding comparisons explicitly include other padding conditions.

Encoder identifiers and feature sites. The recorded model identifiers are `openai/clip-vit-base-patch32`, `openai/clip-vit-base-patch16`, `google/`

siglip2-base-patch16-224, facebook/dinov2-base, and facebook/dinov3-vitb16-pretrain-lvd1689m. Their weights are frozen; probes and learned readouts are fitted separately. The token sites are hidden states 3 and 12. Recorded final feature widths are 512 for CLIP and 768 for SigLIP2 and DINO. The E016 configurations set drop_cls=true for CLIP-B/16 and DINOv2-B and false for SigLIP2-B; E018 sets it to true. The summaries do not specify all normalization, register-token, and pooling operations, so the exact extraction code is needed to interpret the deployed feature consistently across architectures.

Readout configurations. E015 records native token grids of 7×7 (CLIP-B/32) and 16×16 (DINOv2-B), with nonlinear and DFT readouts evaluated on a subset of layers. Its nonlinear readouts share hidden width 64, learning rate 10^{-3}, weight decay 10^{-4}, at most 300 epochs, patience 25, and validation fraction 0.2. E012/E014b/E016/E018 record the same MLP hyperparameters. Most later runs report five seeds; E017/E017b extend to nine seeds and five folds. Seed count does not by itself imply independent data splits: E016/E018 record a fixed base split and vary probe initialization and validation seeds.

Parameter counts constrain but do not uniquely identify architectures. For example, DINOv2-B's positional head has $49\,153 = 64 \cdot 768 + 1$ parameters, despite the recorded 16×16 native grid. This could reflect token selection, spatial reduction, or another fixed preprocessing step. The summary does not establish which. Likewise, $6\,913 = 9 \cdot 768 + 1$ in the DFT arm does not document its frequency selection or complex-valued feature encoding. Loss functions, activations, complete readout modules, linear solvers, preprocessing fit boundaries, and shuffle timing must be recovered from the experiment code rather than inferred from these counts.

Checksums and provenance. Fifteen of the 29 JSON files contain a script SHA-256 field. E015 records feb02b44..., E016 066a69e0..., E018 0295395d..., and E019 64a49c13...; these identify experiment scripts that are absent from the package. The six E017/E017b summaries contain a code revision but no script hash. Three files contain a prereg_hash field: the two E001 summaries share a full digest, whereas E012 stores an abbreviated value. No corresponding design block is provided. Later summaries report matching substrate fingerprints, including cbf2efb8... for the first 300 items, but no image-ID manifest or generated substrate is included to recompute the comparison. These fields are provenance leads, not independent verification of chronology or execution.

Software and compute. E016/E018 record Python 3.10.12, PyTorch 2.10.0+cu128, and NumPy 1.26.4. Several later configurations explicitly set gpu=false; a CUDA-tagged package version does not imply GPU execution. Other summaries omit device or environment details, so the package cannot independently establish an aggregate compute total. Reporting scripts require Python, NumPy, and Matplotlib; the supplied figures were reported to use Matplotlib 3.9.2. Experiment environment lockfiles and checkpoint revisions are absent.

Materials required for a complete release. A reproducible experiment release should add the exact executables identified above, model and dataset revisions, a dependency lockfile, image and fold manifests, all transformation parameters, and feature-cache checksums. It should include predictions with base-image ID, transform, true label, split, seed, fold, and control assignment, along with the original bootstrap and multiple-testing code. Downstream release also requires the estimators and fit boundaries for label-derived features, permutation-null construction, and feature scaling. Archived design documents with their full hashes and dated history are needed to substantiate any preregistration claim. Until these materials are available, figures and tables can be regenerated, but the underlying experimental findings and interval coverage remain independently unverified.

J What permutation invariance implies for a readout

Our comparison concerns how a readout uses an already computed token field. It does not assume that rotating an image acts as a permutation of the resulting tokens. This distinction is necessary when relating the experiments to group-equivariant networks (Cohen & Welling, 2016) and invariant-readout theory (Farrell et al., 2022).

Linear invariant readouts factor through the mean. Let $z = (z_1, \ldots, z_T)$ with $z_t \in \mathbb{R}^d$, and consider a linear readout

$$f(z) = \sum_{t=1}^{T} W_t z_t + b.$$

If f is invariant to every permutation of the token slots for all input token fields, then $W_t = W$ for every t, and hence

$$f(z) = T\, W\, \mathrm{GAP}(z) + b.$$

To see this, exchange slots s and t in an input whose only nonzero token is an arbitrary vector v in slot s. Invariance gives $W_s v = W_t v$ for all v, so $W_s = W_t$. Applying the argument to every pair of slots proves the result. Thus an invariant linear classifier cannot exploit spatial arrangements that have the same token mean. This is a property of the function class, not a statement that a particular label is absent from the token multiset.

Token averaging is not necessarily rotation averaging. For a finite group G with a specified linear action ρ, its Reynolds operator is $|G|^{-1} \sum_{g \in G} \rho(g)$. Averaging a token field over all slot permutations replaces each slot by the same token mean. The experiments do not establish the additional assumptions needed to identify this operator with averaging over image rotations: an action on the learned representation and an equivariance relation for the encoder. Interpolation, image boundaries, and learned positional representations make that identification nontrivial.

Even for an equivariant token field, pooling need not remove the transformation law. Suppose the action permutes the sampled positions and $z_g(p) = \rho(g) z(g^{-1} p)$. Reindexing the sum gives

$$\mathrm{GAP}(z_g) = \rho(g)\, \mathrm{GAP}(z).$$

This identity establishes equivariance of the mean. It does not establish preservation of all information: the mean can vanish, and averaging is generally non-injective. Conversely, it does not imply that every equivariant component is erased. The effect of pooling depends on the representation and the statistic one wants to recover. Retaining a semantic label while poorly decoding rotation sign therefore does not, by itself, identify a Reynolds-projection mechanism.

What the experiments establish. CLIP-B/32 supports much better rotation-sign decoding with the tested positional readout than with the tested invariant readouts. DINOv2-B supports strong decoding with nonlinear invariant readouts as well. These observations motivate different descriptions of the two encoders; neither is an impossibility result for a broader readout family. DeepSets (Zaheer et al., 2017) supplies a family of invariant functions, but a finite network's failed optimization is not a bound on all functions in that family. Similarly, the coordinatewise order statistics in Appendix F discard cross-coordinate token associations and are compressed before classification; they are not a lossless representation of the token multiset. The linear-capacity results of Farrell et al. (2022) provide related theoretical context, but their assumptions do not turn these empirical probe comparisons into a proof of information erasure.