REVEALING INVISIBLE DEFECTS IN VISUAL REPRESENTATIONS: A PSYCHOMETRIC INSTRUMENT FOR BINDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Task accuracy can conceal differences in how visual representations respond to changes in attribute–object pairings. We introduce BInD, the Binding Intervention Difference, a diagnostic protocol for frozen visual encoders. BInD compares cosine-distance responses to pairing-preserving and pairing-breaking image edits, without human annotation or a trained probe at scoring time. It measures readout-specific sensitivity, not the total binding information recoverable by an arbitrary decoder. Synthetic and real-photo suites cover color, load, arrangement, and a size control at pooled and patch interfaces. Across 32 encoders, we report reliability, axis-specific associations, and comparisons with eleven intrinsic metrics under training-objective and capability controls. Within-recipe interventions change several metrics substantially without consistently improving binding readouts. Our strongest application is compression: two DINOv2 merge configurations have nearly identical ImageNet accuracy but retain 86% versus 34% of the reference color score. Pruning can preserve that score while degrading segmentation. Distillation and vision-language case studies further show why the interface and evaluation task matter. We provide suite specifications, a scoring pipeline, and an interpretation guide with explicit limits on causal and real-image claims.

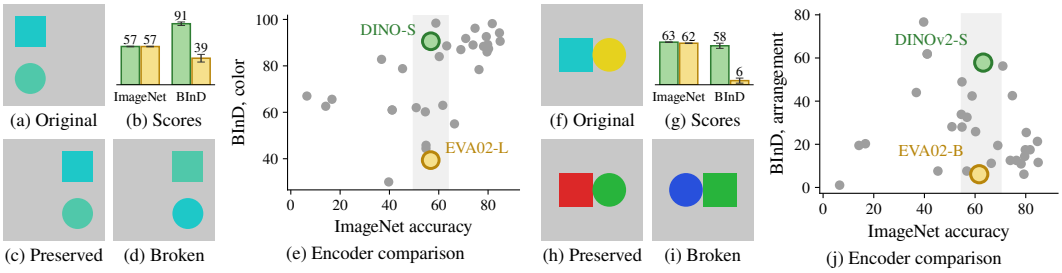

Figure 1: **Similar classification accuracy, different BInD profiles.** Color in (a)–(e) and arrangement in (f)–(j). Bars report ImageNet kNN and BInD forced-choice accuracies in percent; scatter plots show all 32 encoders.

1 INTRODUCTION

Visual encoders are commonly compared by downstream accuracy (Kornblith et al., 2019; He et al., 2022; Oquab et al., 2024). These evaluations measure task utility, but do not identify which structural properties differ between equally accurate representations (Hooker et al., 2019; Geirhos et al., 2020). Figure 1 illustrates this distinction: encoders with similar ImageNet scores respond differently when attribute–object pairings change.

We study feature binding: which attribute belongs to which object (Treisman & Gelade, 1980; Treisman & Schmidt, 1982). Our aim is an inexpensive diagnostic that reads an encoder directly and reports a profile rather than a single quality ranking. Intrinsic metrics provide direct, label-free summaries but need not distinguish particular conjunctions (Garrido et al., 2023; Chung & Kim, 2026). Compositional benchmarks reveal specific failures, often through text alignment or an entire

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.

vision-language model (Yuksekgonul et al., 2023; Hsieh et al., 2023; Tong et al., 2024). These measurements are complementary, not interchangeable.

The practical question is not only whether an encoder succeeds on a task, but which conjunctions remain accessible when its representation changes. Compression can leave classification nearly unchanged, while distillation can transfer a teacher's weakness together with its strengths. A profile is useful precisely when one scalar cannot distinguish these outcomes: it identifies a property to inspect without replacing evaluation on the task the system must perform.

BInD, the **B**inding **In**tervention **D**ifference, moves the diagnostic intervention to the image. For example, a pairing-breaking edit exchanges the colors of a circle and a square while retaining their attribute inventory. A pairing-preserving edit supplies a contrast with controlled appearance changes. The difference between their cosine-distance responses measures the preference for preserving the pairing under this stimulus design. It does not establish that a low-scoring representation contains no decodable binding information.

Our contributions are threefold. **Instrument:** a frozen-encoder protocol with four axes, two interfaces, curated synthetic and real-photo suites, and scoring code. **Measurement evidence:** reliability, cross-axis and controlled metric comparisons, and within-recipe metric interventions, with the distinction between internal consistency and independent construct validity made explicit. **Diagnostic applications:** compression configurations with similar task scores but different BInD profiles, followed by distillation, patch-interface, and vision-language case studies. Compression provides the central result; the other applications define the protocol's scope and limits.

2 Related Work

Representation evaluation. Task evaluations measure classification, segmentation, or depth (Deng et al., 2009; Zhou et al., 2019; Ignatov et al., 2024). Intrinsic alternatives summarize spectra, distributions, dimension, or local sensitivity (Roy & Vetterli, 2007; Garrido et al., 2023; Wang & Isola, 2020; Thilak et al., 2024; Chung & Kim, 2026). Diagnostic benchmarks instead test named visual or compositional weaknesses (Tong et al., 2024; Yuksekgonul et al., 2023; Hsieh et al., 2023; Thrush et al., 2022; Huang et al., 2023; Ghosh et al., 2023). We compare eleven intrinsic metrics on our color target; this does not assess every metric against its original purpose. Appendix A gives the full comparison.

Binding and direct encoder tests. Object-centric architectures address binding through structured representations (Greff et al., 2020; Locatello et al., 2020; Seitzer et al., 2023). Measurement studies examine CLIP concepts (Lewis et al., 2024; Koishigarina et al., 2026), object binding in ViTs (Li et al., 2025b), and binding mechanisms in embedding models (Uselis et al., 2026). Huang et al. (2026) formalize binding information and evaluate it by decoding latent representations. Chung & Kim (2026) also use frozen-encoder cosine tests on synthetic shape–position binding. Thus neither synthetic stimuli nor a frozen cosine readout alone is new. BInD contributes the preserved/broken contrast, a multi-axis instrument, curation and interpretation rules, and diagnostic interventions. An independent decoder-based comparison remains to be performed.

Intervention-based analysis. Control probes and representation interventions test what an embedding supports (Hewitt & Liang, 2019; Ravfogel et al., 2020; Elazar et al., 2021; Belinkov, 2022). BInD intervenes on the stimulus and uses no trained decoder at scoring time. Construction controls the intended pairing change; it does not guarantee equal nuisance responses in every nonlinear encoder.

3 BInD

3.1 Definition

Let $x \in \mathcal{X}$ be an image, $x_{\text{preserved}}$ a pairing-preserving edit, and x_{broken} a pairing-breaking edit. For an encoder f and cosine distance $d(u, v) = 1 - \cos(u, v)$, define

$$\text{BInD}(x) = d\big(f(x), f(x_{\text{broken}})\big) - d\big(f(x), f(x_{\text{preserved}})\big). \tag{1}$$

A positive margin means that the preserved edit is closer to the original under this readout. A zero or negative margin means that this preference is absent or reversed; it is not proof of information loss. BInD is therefore a measure of *readout-specific binding sensitivity* under the prescribed edits. Matching pixel distances helps control edit magnitude but does not identify an encoder-independent binding quantity.

The preserved edit is essential to this interpretation: distance from a broken edit alone can also increase because an encoder responds strongly to any appearance change. Subtracting the preserved response provides a matched contrast, rather than treating every large response as evidence of binding. Section 4.4 compares this difference with its components and a ratio.

The *global* interface uses pooled embeddings. The *dense* interface compares corresponding patch-grid positions and averages:

$$\mathrm{BInD}_{\mathrm{dense}}(x) = \frac{1}{P}\sum_{p=1}^{P}\left[d\big(f_p(x), f_p(x_{\mathrm{broken}})\big) - d\big(f_p(x), f_p(x_{\mathrm{preserved}})\big)\right]. \tag{2}$$

This requires aligned grids; token-reduced VLMs use a separate pooled readout. At either interface the margin is continuous. Forced-choice accuracy instead selects $\hat{c} = \arg\max_c \cos(f(x), f(c))$ and is chance-corrected as $(a - c)/(1 - c)$, where a is raw accuracy and c is chance. For the same two candidates, correctness is equivalent to a positive margin, apart from ties. Our margin and accuracy suites use separate stimuli, but share this measurement principle.

3.2 Axes

The axes follow attribute-binding and spatial tests (Treisman & Gelade, 1980; Treisman & Schmidt, 1982; Ghosh et al., 2023; Huang et al., 2023). **Color** binds colors to two shapes; **load** uses three objects and tests the added conjunction demand. **Arrangement** binds objects to positions. **Size** asks which object is larger and serves as an empirical capability control because it tracks classification in our comparisons. This role does not make size a universal measure of general capability. We measure static representations, not the attentional process proposed in human binding theories.

Color and load are related by design, but increasing the number of objects also increases the conjunctions that must be kept distinct. Arrangement changes a different relation, the assignment of objects to places. Keeping these tests separate allows an encoder to perform well on color assignment yet poorly on layout, instead of averaging that contrast away.

3.3 Suites

Candidates share an attribute inventory and vary its pairing. Color, load, and size use two candidates; arrangement uses four. A master seed fixes generation. Curation screens pixel-cosine shortcuts, balances answer indices, removes repeated signatures, and stratifies attributes. Item selection does not use encoder responses. Both raw-pixel and random-projection cosine baselines score at chance after curation; this excludes those shortcuts, not every possible low-level strategy.

The four curation rules address different failure modes: hygiene screens a specified shortcut, balance prevents an answer-position preference, deduplication avoids repeated conjunction signatures, and stratification limits uneven attribute coverage. Thus item quality is defined by the construction and control examinees, rather than by which encoder answers correctly.

The frozen suite contains 1,000 items per axis except load, whose finite near-hue factor space yields 724. Color and load use similar hues because distinct colors put 19 of 32 encoders above .95. This difficulty choice used the observed ceiling, whereas subsequent item filtering did not use model scores. Separate margin suites contain 300 triplets per axis. Their broken and preserved pixel distances differ by at most .005 on the binding axes. A one-object control separates the same hues at .82–1.00 across encoders. Its association with the color axis is only $+.24$ and is not significant. This argues against hue discriminability alone explaining the two-object scores. The easy distinct-color variant remains available as a ceiling check, rather than being combined with the harder axis. Appendices C and E give the curation and controls.

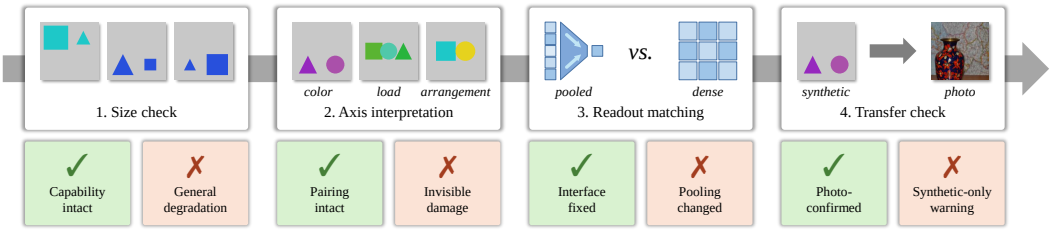

Figure 2: **Interpretation guide, not a causal decision tree.** Size is read first, because a decline there makes a binding-specific attribution ambiguous; color, load, and arrangement are then read separately; the readout is held fixed across a comparison; and photographs are tested before a real-image claim. The stimuli are items from the released suite, and a profile suggests follow-up tests rather than identifying a unique failure mechanism.

3.4 Interpretation

Read the four axes separately (Figure 2). A size decline makes a binding-specific attribution ambiguous; it does not show that binding is unaffected. When size is stable, a color or load decline is a candidate selective change. Arrangement reports sensitivity to layout, not the presence or absence of all structure. Appendix B links observed patterns to tested interventions, not uniquely inferred causes.

For cross-encoder comparisons, use chance-corrected accuracy rather than the capability-sensitive margin. For within-encoder comparisons, margins can resolve finer changes, but capability and nuisance sensitivity must still be checked. Compression uses reference-normalized accuracy (Section 5.1). Keep interfaces fixed, and rescore deployment comparisons at each model's native resolution (Appendix P). Close-family rankings and individual training-run differences require additional uncertainty analysis.

The shared-resolution survey uses each encoder's preprocessing at 224. Native-resolution comparisons answer a different deployment question: for DINOv2 and DINOv3, a large load gap at 224 becomes a much smaller difference at their respective native resolutions. The resolution and readout therefore belong to the reported profile, not merely to implementation details.

3.5 Release and real-photo scope

The release comprises item specifications, generation and scoring code, checksums, and an interpretation guide. The synthetic suite is primary because its factors and tested shortcuts are controlled. A real-photo suite composites object cutouts into photographs, with 400 items per axis. Synthetic–photo rank transfer is significant for size and load but not color ($r = +.30$); photographic arrangement is also solvable by pixel cosine (.99). We therefore do not claim universal real-image rankings or use photographic arrangement as clean construct validation. Photo interventions contextualize synthetic changes, whose magnitude is not directly transferable (Appendix M).

The significant rank-transfer coefficients are $+.77$ for size and $+.66$ for load. Hybrid stimuli using photographic silhouettes with flat colors, or synthetic shapes with photographic color texture, improve color agreement to $+.67$–$+.68$. These controls implicate stimulus material in the transfer gap without establishing that all suites measure an identical latent construct.

4 Validation

We organize evidence around reliability, discriminant and incremental validity (Campbell & Fiske, 1959; Sechrest, 1963; Cronbach & Meehl, 1955). Split-half reliability is .99 per axis (Appendix C). The reported standard error of a single encoder's axis score is at most .018. Reliability concerns repeatability over items; the cross-axis tests ask whether the profile is redundant, and the metric comparisons ask whether associations remain after specific alternative explanations are controlled. The following tests establish consistency and limited predictive associations, not a complete validation of latent binding information.

Margin	Color	Load	Arr.
Color	$+.\mathbf{56}$	$+.39$	$-.28$
Load	$+.37$	$+.\mathbf{63}$	$-.16$
Arr.	$-.37$	$-.10$	$+.\mathbf{63}$

(a) Cross-axis associations

Metric	Obj.	+cap.	Metric	Obj.	+cap.	Metric	Obj.	+cap.
PR	$+.51$	$+.29$	Alpha-ReQ	$-.30$	$+.30$	GS	$-.52$	$-.31$
Eff. rank	$+.45$	$+.20$	Uniformity	$-.19$	$+.41$	TokenEnt	$-.65$	$-.\mathbf{57}$
Isotropy	$+.13$	$-.38$	TwoNN	$-.33$	$+.16$	JER	$+.11$	$-.12$
RankMe	$+.25$	$+.12$	LiDAR	$+.28$	$+.06$	BInD	$+.67$	$+.\mathbf{56}$

(b) Metric comparisons

Table 1: **Controlled associations across 32 encoders.** (a) Margin–accuracy associations under objective and capability controls; bold diagonals meet the reported significance criterion. (b) Color-target partial correlations: objective only (obj.) and objective plus ImageNet kNN (+cap.). Bold +cap. entries survive Benjamini–Hochberg correction over twelve tests.

4.1 The axes are related but nonredundant

Color and load correlate most strongly ($+.68$); arrangement correlates weakly with them ($-.21$ and $-.16$). With training objective and ImageNet kNN controlled, each binding-axis margin associates most clearly with its own accuracy (Table 1a). The off-diagonal results are not proof of independence: color–load associations remain moderate. These findings support separate reporting rather than an aggregate binding rank. Because margin and accuracy share the intervention principle, their diagonal associations are evidence of internal consistency, not independent convergent validity. One association does not share that principle. The dense reading predicts segmentation fragmentation and depth across 31 encoders under objective controls (Section 5.3), where the targets are supervised probe scores rather than intervention responses. A designed-representation comparison is reported in Appendix E.

4.2 What intrinsic metrics predict under controls

We compare eleven metrics spanning spectra, distributions, intrinsic dimension, augmentation discrimination, and sensitivity (Appendix F). For each, we regress both the metric and color accuracy on objective indicators and then add ImageNet kNN.

Objective indicators account for the fact that a training recipe can move both quantities together. The capability control asks whether a metric adds information beyond a conventional classification score. The frozen set spans six objective groups; the capability gallery contains ten images per class, with five queries per class (Appendix D). The twelve-test family includes BInD and uses Benjamini–Hochberg correction at $q = .05$ (Benjamini & Hochberg, 1995).

Under both controls, BInD has a positive association ($+.56$) and TokenEnt a negative one ($-.57$); both survive correction (Table 1b). *A negative association is predictive*, but does not support a "more diversity, better binding" interpretation. The remaining metrics do not meet the corrected criterion here. This is not a general rejection of their intended uses, nor does a nonsignificant coefficient establish no relationship. The color-target comparison cannot resolve all claims about spatial binding.

Controls can reveal as well as reduce an association. For example, TokenEnt changes from $-.11$ without controls to $-.65$ with the objective controlled, whereas PR changes from $+.58$ without controls to $+.29$ under both controls. The contrast motivates reporting the controlled and uncontrolled readings rather than treating either alone as a binding-specific result.

A larger capability gallery, aggregation to 17 training recipes, and alternative objective assignments preserve the main pattern (Appendix D). The larger gallery ranks encoders almost identically (Spearman .997), with BInD and TokenEnt partials of $+.58$ and $-.58$. Aggregating size variants to one point per recipe gives $+.66$ and $-.83$, respectively. These checks reduce specific confounds but do not make model families independent or establish causal effects. BInD's same-axis margin–accuracy association is not an upper bound on another metric's predictive performance.

4.3 Metric optimization is not sufficient for improvement

Within a fixed ViT-S recipe, we add a differentiable metric term to 200 steps of supervised continuation. Each counted arm uses four seeds. Five metrics move in both directions; uniformity

is evaluated in one direction, and isotropy is excluded when classification collapses. The changes include JER at $2.0\times$ /$0.07\times$, PR at $1.5\times$ /$0.01\times$, and TokenEnt at $2.1\times$ /$0.07\times$ for the upward/downward arms. Appendix G gives all factors and exclusion rules. For example, PR falls to 1.6 while classification remains $.775$. Counted-arm mean binding changes do not exceed the control range's width ($.45$–$.53$), while within-arm seed differences reach nine points (per-seed values in Appendix G). These descriptive comparisons are not equivalence tests.

The principal comparisons also require a functioning classification anchor: counted arms retain top-1 accuracy between $.70$ and $.80$, whereas the excluded isotropy descent reaches $.08$. Spectral metrics can move together under these interventions, so targeting one regularizer does not isolate every other geometric property.

The principal battery is a distinct-color four-way task, not the frozen near-hue two-way task. For TokenEnt descent, the added near-hue check scores $.86$–$.91$, versus $.90$–$.95$ for controls, and shows no gain. A margin-targeted arm can reduce four-way performance to chance without reducing near-hue performance, which limits transfer between the readouts. Under a second, contrastive anchor, four capability-preserving arms also fall within a broad control range ($.48$–$.79$). The results show no consistent binding improvement from these metric optimizations; they neither prove zero effect nor invalidate a purely observational predictor.

4.4 Estimator and readout choices

Under both controls, the difference correlates with color accuracy at $+.56$, the broken response at $+.54$, the ratio at $+.50$, and the preserved response at $+.22$. This does not establish a significant predictive advantage for the difference. We choose it for the preserved/broken contrast and to avoid small ratio denominators. Pixel matching controls edit size only at that baseline. The margin's capability association changes across settings ($-.83$ within one family and $+.65$ across objectives); accuracy's corresponding within-family association is $-.16$, nonsignificant. This motivates, but does not universally guarantee, the accuracy-based comparison rule (Appendices E and O).

5 Diagnostics

We first test compression, where the manipulated design element is known. The subsequent case studies are narrower and do not identify causes from a profile alone. Throughout, "preservation" refers to a BInD score, not a fraction of all recoverable binding information.

5.1 Equal task accuracy can hide different compression profiles

For raw accuracy a, uncompressed accuracy a_0, and chance c, report

$$\text{Preservation}(\%) = 100\,\frac{a-c}{a_0-c}. \tag{3}$$

The reference must exceed chance; values can be negative or greater than 100%. We compare token reduction at matched settings (Bolya et al., 2023; Liang et al., 2022). Here r denotes the experiment's token-reduction rate, not a common percentage across architectures.

Configuration	Color (%)	ImageNet kNN ↑	ADE20K mIoU ↑	NYUv2 RMSE ↓
Uncompressed	100	.785	.316	—
ToMe, no prop. attention	34	.7643	.280	.766
ToMe, official prop. attention	**86**	.7642	.285	.765
CLS-saliency pruning	100	.768	.184	—

Table 2: **The central compression comparison: DINOv2 at $r = 16$.** Two merge settings have nearly identical task scores but different color preservation. Task columns are raw scores; only color is reference-normalized by Equation 3. A dash means that the value is not reported here. Protocols and additional comparisons are in Appendix H.

Result 1: proportional attention changes the color score. On DINOv2, restoring token-size weighting raises color preservation from 34% to 86%, while ImageNet kNN is effectively unchanged (Table 2). This weighting is ToMe's existing *proportional attention*, not a new merging method (Bolya et al., 2023). Our contribution is the diagnostic contrast between its presence and absence.

A merged token may represent many original patches. Without size weighting, attention treats that token as one contributor regardless of the mass it represents. The comparison tests this existing design element while keeping the encoder and reduction setting fixed, rather than contrasting two unrelated compression pipelines.

Result 2: saliency choice changes pruning's profile. Table 3 makes the cross-model comparison explicit. Feature-norm pruning preserves 76–89% of the color score on four non-RoPE encoders; random pruning preserves −2–18%. Attention-based choices vary more across these models. On RoPE-based DINOv3, random and norm pruning both retain 56%, so the pattern does not generalize without qualification.

	Merging			Pruning			
Encoder	No prop.	Official	PiToMe	EViT	Zero-TP	Norm	Random
DINOv2	34	86	63	100	89	88	16
CLIP	−4	82	56	85	71	89	1
DeiT3	69	97	82	63	23	88	18
MAE	24	79	28	45	27	76	−2

Table 3: **Color preservation depends on the design element.** Percent of the uncompressed score at $r = 16$; ToMe's official proportional-attention configuration is shown alongside its ablation. PiToMe (Tran et al., 2024), EViT (Liang et al., 2022), and Zero-TPrune (Wang et al., 2024) retain their reported settings. Appendix H gives the matched ToMe and saliency comparisons.

Result 3: color preservation does not imply dense-task preservation. CLS-saliency pruning retains the DINOv2 color score and changes kNN from .785 to .768, yet ADE mIoU falls from .316 to .184 (Table 2). Thus classification, BInD, and segmentation supply different information; none replaces the others.

Mechanism and scope. In the DINOv2 $r = 16$ inspection, none of 2,112 merges crosses two different objects, and forbidding such merges does not recover the color score. This rules out that particular explanation in the inspected setting, not every mixing-related mechanism. Proportional attention retains at least 79% on the four non-RoPE encoders at the tested $r \leq 16$ settings; recovery weakens at larger reductions and differs on DINOv3 (Appendix H). The intervention supports a setting-specific design conclusion, not a universally safe compression rule.

5.2 Distillation objectives transfer different profiles

A supervised ViT-S student is distilled for 5,000 steps under four alignment objectives (Romero et al., 2015; Park et al., 2019; Yang et al., 2024; Sun et al., 2024). With a DeiT3 teacher, arrangement starts at .42 in the student and .32 in the teacher. Distance alignment lowers the student's advantage; logit alignment preserves it along all eleven checkpoints. Relation alignment improves arrangement, while standardized logits vary widely across seeds (Table 4). A ConvNeXt teacher reproduces the qualitative tendency, but distance alignment reaches the teacher level in only one of two seeds. These small comparisons motivate profile-aware evaluation, not a universal ranking of distillation objectives (Appendix I).

The teacher contrast provides a separate check. An I-JEPA teacher scores .71 on color against the student's initial .94; the distilled student ends at .73 and .78 on two seeds. With a color-strong DINOv2 teacher, the student ends at .77 and .86. Thus the choice of teacher and the quantity aligned both matter to what the student retains.

For the second objective comparison, relation alignment reaches .49 on both seeds, whereas distance alignment ends at .33 on one and .44 on the other. The trajectories therefore qualify the endpoint

ranges: a tendency that recurs across teacher pairs need not have the same magnitude or hold on every seed.

Alignment	Distances	Logits	Standardized logits	Relations
Arrangement	.34–.37	.43–.44	.32–.58	.47–.51

Table 4: **Distillation changes depend on the alignment objective.** Final arrangement accuracy across seeds for the DeiT3-teacher comparison; student initialization .42, teacher .32. Two seeds per arm, except four for standardized logits. These ranges are not confidence intervals.

5.3 Patch and pooled interfaces answer different questions

Dense tasks consume a patch grid rather than only a pooled vector (Darcet et al., 2024; Oquab et al., 2024). Equation 2 provides a corresponding intervention readout. Appendix J compares it with ADE20K segmentation, boundaries, fragmentation, and NYUv2 depth across 31 encoders; one is excluded for a diverging probe. These targets distinguish semantic recognition from geometric organization even within one dataset. Controlling segmentation leaves the depth association at $+.65$; the dense reading carries depth information beyond segmentation quality.

The global size axis provides a complementary result: its classification associations range from $+.45$ to $+.71$ over eight tasks, with five surviving correction. The tasks include ImageNet variants, Pets, Flowers, Food, Aircraft, and iNaturalist 2019.

These results do not establish an exact three-part decomposition of representation quality, and the supervised downstream evaluations assess the utility of an annotation-free diagnostic without making the validation targets label-free. The practical implication is to specify the axis and interface for each target and evaluate dense-task utility directly.

5.4 VLM readouts differ in sensitivity to compression

We compare BInD on the mean of vision-tower tokens with two-alternative caption-likelihood evaluation and dual-encoder retrieval (Liu et al., 2024; Li et al., 2025a; Hsieh et al., 2023). The language model consumes individual tokens, not that mean. The caption-likelihood task is not open-ended generative VQA.

The systems are LLaVA-1.5-7B, with a CLIP ViT-L/14-336 tower, and LLaVA-Interleave-0.5B, with SigLIP so400m. The photographic caption and retrieval tests use 666 attribute-swap items from SugarCrepe. At $r = 24$, their towers retain 24 of 576 and 105 of 729 tokens respectively, so the same r is not an equal compression fraction across systems.

At $r = 8$ and $r = 18$, color preservation is 66% and 62%, while size remains at 97% and 101% (Appendix L). These are candidate selective changes under the synthetic readout. At moderate reduction, however, the real-photo suite does not confirm a comparable loss. We treat the synthetic result as an early warning, not demonstrated real-image failure.

Readout	No prop. attention	Official ToMe
Synthetic color BInD	95%	95%
Photo paired-composite test	60%	21%
Two-alternative caption likelihood	29%	20%
Dual-encoder retrieval	12%	n.s.

Table 5: **Reported relative decreases at extreme compression ($r = 24$).** Each row uses its own task scale and reference; the percentages are not calibrated measurements of one common loss. The extreme synthetic merge run also retains only 16% on size (Appendix L), so that run's 95% color decrease is not binding-specific evidence. "n.s." denotes no significant decrease, not zero change.

Extreme compression produces larger changes across several readouts (Table 5), but also broader degradation. In raw terms, the synthetic color reading falls from .85 to .52 against a .50 chance floor, while caption likelihood falls from .82 to .73 (Appendix L). Different stimulus sets, baselines, and chance levels prevent interpreting their decrease ratios as universal "hiding factors." Weighting the surviving tokens by their patch mass does not recover the reading (Appendix L), so the decline is not

an artifact of the uniform mean; a token-aware control through the language model remains open. The present evidence supports using multiple readouts, not a claim that deployment benchmarks systematically miss a proven encoder defect.

A separate ViT-B/16 retrieval experiment supplies a narrower behavioral check. At $r = 16$, attribute-swap performance declines 1.6–3.0 times as fast as its object-replacement control; merging is the significant comparison ($p = .009$). This supports an attribute-sensitive effect in that retrieval setting, without calibrating all of the VLM readouts to a common damage scale.

5.5 Training recipes have different empirical profiles

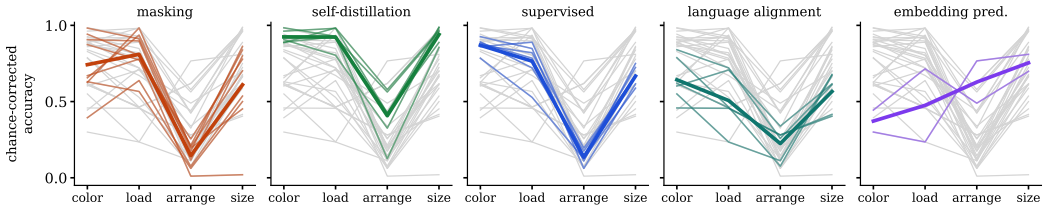

Figure 3: **Profiles differ across training-recipe groups.** All encoders are grey; the highlighted group is colored and its mean is bold. These observational groups include differences in losses, augmentations, and training data.

The groups in Figure 3 differ in their synthetic profiles. For example, I-JEPA-G exceeds DINOv2-B on arrangement at the shared 224 interface (.825 versus .672), but not on color. Five additional architecture families reproduce the broad supervised pattern, and a 13-encoder ImageNet-1k subset retains recipe differences (Appendix K). These controls do not isolate the loss function from the rest of training.

The additional families cover window attention, MLP-based models, a gated convolutional network, and a state-space hybrid. All five exhibit the reported supervised pattern: relatively high color and load scores, lower arrangement, and a negative arrangement margin. Within the ImageNet-1k subset, recipe differences persist despite sharing the training-image source. Language-aligned models remain confounded with their broader training data, so "objective" here names the training recipe rather than an isolated loss term.

For MMVP (Tong et al., 2024), we measure embedding discrimination on 135 photograph pairs in nine categories, normalized by random-pair distances within each category. The main analysis excludes CLIP and DINOv2, which were involved in selecting the pairs, leaving 25 encoders. Ten corrected associations involve arrangement or size, and nine hold on a held-out half of the pairs (Appendix N). Even MMVP's Color category associates with size ($+.83$), not the color axis ($-.07$), illustrating that a benchmark category and a diagnostic construct need not coincide. These embedding-level associations do not establish that spatial layout uniquely causes MMVP failures or that attribute binding is irrelevant.

6 Conclusion

BInD measures how frozen visual encoders respond to controlled pairing edits at specified interfaces. Its central contribution is a diagnostic profile that complements task accuracy. The compression experiments show why this matters: nearly identical classification scores can coexist with substantially different color responses, and preserved color responses can coexist with poorer segmentation. The released protocol makes these distinctions inspectable without a trained scoring probe.

Limitations. BInD is readout-dependent and does not measure all decodable binding information. Decoder-based construct validation on learned encoders remains untested; a designed-representation check with a known answer is reported instead. Metric interventions have limited seeds and do not establish equivalence to zero effect. Synthetic color rankings and moderate-compression changes are not confirmed generally on photographs. VLM pooling, limited distillation pairs, and exploratory dense statistics restrict the secondary claims. Profile patterns suggest experiments; they do not uniquely identify causes.

AI USE STATEMENT

A large language model assisted coding, literature search, and manuscript revision, including organization, readability, and checks on the scope of claims. No new experimental measurements were produced during manuscript revision. The authors are responsible for verifying the code, numerical results, references, and final text; Appendix R provides further detail.

REPRODUCIBILITY STATEMENT

The measure and suites are specified in Section 3; Appendices C, D, and T describe the generation, scoring, controls, and release artifacts.

ETHICS STATEMENT

The synthetic suite is generated; the real-photo release uses provenance and masks rather than redistributed photographs. Appendix S discusses intended use and limitations.

REFERENCES

Kumar K. Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Jiwan Chung and Seon Joo Kim. Global geometry is not enough for vision representations, 2026. arXiv:2602.03282.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, 2021. doi: 10.1109/ICCV48922.2021.00823.

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 (EMNLP)*, 2019.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Lianghuan Huang, Yihao Li, Saeed Salehi, Yingshan Chang, Ansh Soni, and Konrad P. Kording. Formalizing the binding problem. In *International Conference on Machine Learning (ICML)*, 2026.

Dmitry Ignatov, Andrey Ignatov, and Radu Timofte. Virtually enriched nyu depth v2 dataset for monocular depth estimation: Do we need artificial augmentation? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6177–6186, 2024. doi: 10.1109/CVPRW63382.2024.00622.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations (ICLR)*, 2026.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL*, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3d in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P. Kording. Does object binding naturally emerge in large pretrained vision transformers? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024.

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

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1447–1454, 2006.

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

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Prashant C. Raju. Geometric stability: The missing axis of representations. *arXiv preprint arXiv:2601.09173*, 2026.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Lee Sechrest. Incremental validity: A recommendation. *Educational and Psychological Measurement*, 23(1):153–158, 1963.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun 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*, 2026.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning (ICML)*, 2025.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *International Conference on Learning Representations (ICLR)*, 2024.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *Computer Vision – ECCV 2022*, pp. 516–533. Springer, 2022.

Hoai-Chau Tran, Duy M. H. Nguyen, Duy M. Nguyen, Trung-Tin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Y. Zou, Binh T. Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–141, 1982.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

Arnas Uselis, Darina Koishigarina, and Seong Joon Oh. How can embedding models bind concepts? In *International Conference on Machine Learning (ICML)*, 2026.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Hongjie Wang, Bhishma Dedhia, and Niraj K. Jha. Zero-TPrune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16070–16079, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. ViTKD: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1379–1388, 2024.

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.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

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

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

REVEALING INVISIBLE DEFECTS IN VISUAL REPRESENTATIONS: A PSYCHOMETRIC INSTRUMENT FOR BINDING

APPENDIX

A RELATED WORK IN FULL

Downstream evaluation. ImageNet classification and its variants (Deng et al., 2009), transfer suites (Kornblith et al., 2019), segmentation, and depth (Zhou et al., 2019; Ignatov et al., 2024) measure task utility. An aggregate score does not describe every change in a representation (Hooker et al., 2019; Geirhos et al., 2020). It also depends on the dataset and evaluation head. BInD complements these evaluations by measuring responses to controlled edits without human annotation or probe training at scoring time; downstream evaluations used to assess its utility retain their own supervision.

Intrinsic representation metrics. We compare spectral measures (effective rank, RankMe, and participation ratio) (Roy & Vetterli, 2007; Garrido et al., 2023; Gao et al., 2017); distributional measures (isotropy and uniformity) (Mu & Viswanath, 2018; Wang & Isola, 2020); spectral decay (Agrawal et al., 2022); augmentation discrimination (Thilak et al., 2024); intrinsic dimension (Facco et al., 2017); token entropy (Skean et al., 2025); split-half stability (Raju, 2026); and Jacobian sensitivity (Chung & Kim, 2026). These summaries can be useful predictors without uniquely specifying attribute–object pairings. Section 4 examines their associations with our color target under objective and capability controls. TokenEnt retains a negative association; most others do not meet the corrected criterion. These findings do not establish that intrinsic geometry can never predict binding, and the optimization experiments do not substitute for tests of observational prediction.

Diagnostic and compositional benchmarks. MMVP (Tong et al., 2024), image–text composition tests (Yuksekgonul et al., 2023; Hsieh et al., 2023; Thrush et al., 2022), and generation-side benchmarks (Huang et al., 2023; Ghosh et al., 2023) distinguish named weaknesses. Their text- or system-level readouts address different questions from a direct visual-encoder test. Section 5.5 relates BInD axes to MMVP embedding discrimination, without attributing a unique cause to each benchmark category.

Closest binding measurements. Binding has been studied through object-centric representations (Greff et al., 2020; Locatello et al., 2020; Seitzer et al., 2023), concept probes (Lewis et al., 2024), unimodal versus cross-modal tests (Koishigarina et al., 2026), pretrained ViT analyses (Li et al., 2025b), and embedding-model mechanisms (Uselis et al., 2026). Huang et al. (2026) study binding information using latent decoding, while Chung & Kim (2026) include frozen cosine tests of synthetic shape–position binding. BInD's distinction is the preserved/broken intervention contrast integrated with multiple axes, curation rules, interface-specific readouts, and compression diagnostics. An independent decoding comparison would test whether BInD tracks recoverable binding beyond its own cosine geometry; the present experiments do not provide that validation.

B AN EMPIRICAL INTERPRETATION GUIDE

Table 6 links patterns observed in our experiments to candidate explanations and follow-up comparisons. It is not an inverse model that identifies a cause from a profile alone. Different mechanisms may produce the same pattern, and a stable size score does not exclude every general-capability confound. The interventions and their scope are reported in Appendices H and I.

At the tested $r \leq 16$ settings, proportional attention preserves at least 79% of the color score on the four non-RoPE encoders. This operating range does not cover the RoPE exception or guarantee preservation at extreme reduction.

Observed profile	Candidate explanation	Follow-up tested here
Size declines with other axes	Broader degradation may coexist with binding changes	Do not attribute the decline specifically to binding without further controls.
Color/load decline; arrangement is stable	In some merge settings, attention omits token mass	Compare official proportional attention with its ablation.
Color/load decline only on some models	Saliency may be model-dependent	Compare CLS attention, feature norm, and random selection.
Arrangement declines; color/load are stable	Distance alignment can transfer a teacher's weak axis	Compare distance, logit, and relation alignment on the same pair.
Patch and pooled readings differ	Interface-specific sensitivity or pooling changes	Evaluate aligned patch readouts and the relevant dense tasks.

Table 6: **Profiles suggest follow-up tests, not unique causes.** The second and third columns describe explanations and comparisons examined in this paper, subject to the stated model and reduction ranges.

C THE RELEASE

Generation and curation. Every item comes from one master seed. We generate up to ten thousand items per axis and then filter with rules that never consult a model score, since a suite selected by scores would mirror the models it is meant to test. The rules are a pixel screen, a balance and trim on the index of the correct answer, deduplication of repeated attribute signatures, and stratified sampling over colors, shapes, positions, and counts.

released file	axis in the paper	items	pixel-only examinee
`color2h`	color	1,000	0.500
`load3h`	load	724	0.500
`size`	size	1,000	0.500
`arrange`	arrangement	1,000	0.250
`color2`, `load3`	easy variants, not scored	1,000 each	0.500

Table 7: The frozen suite. A pixel-only examinee scores at chance on every axis.

Why load holds 724 items. The near-hue factor space for three objects is finite, at 864 unique signatures, and the screen and trim leave 724. It is a property of the design rather than a shortfall.

Reliability. Splitting each axis's items in half and correlating the encoders' half-scores gives a Spearman–Brown reliability of .99 on every axis. The reported standard error of a single encoder's axis score is at most .018. These summaries do not make every few-point pairwise difference reliable; close comparisons require uncertainty for the difference, accounting for shared items and model dependence.

Continuity across the freeze. Ranking the 32 encoders on the pre-freeze 400-item scoring against the frozen 1,000-item scoring agrees at 0.91 or above on all six released tasks, the four axes and the two easy variants, and the four margin streams agree at 1.000, since those stimuli were carried over unchanged.

What ships. Item specifications rather than pixels, the generation code, the scoring pipeline, the numbers behind the paper as a CSV, SHA256 checksums over both specifications and code, pinned library versions, and a reproduction README. The same code rebuilds the same suite without a single image changing hands.

D EXPERIMENTAL SETUP

Scoring and resolution. Each encoder uses its own preprocessing at the shared 224 interface. Normalized embeddings are compared by cosine similarity. No head is trained for BInD scoring; downstream task evaluations use their stated evaluation protocols. Rescoring the four DINOv2

objective	n	encoders
masking	11	BEiT-B/L, BEiTv2-B/L, EVA02-B/L, I-JEPA-G/H, MAE-B/L/H
language alignment	6	CLIP-B16/B32/L14, EVA-CLIP-B, MetaCLIP-B (two variants)
supervised	6	ViT-B, ViT-L, DeiT3-S, DeiT3-B, ResNet50, ConvNeXt-B
self-distillation	6	DINO-S/B, DINOv2-S/B/L/G
variance-decorrelation	2	Barlow Twins, VICReg
clustering	1	SwAV

Table 8: The frozen set of 32 encoders (He et al., 2022; Bao et al., 2022; Peng et al., 2022; Fang et al., 2024; Assran et al., 2023; Caron et al., 2021; Oquab et al., 2024; Radford et al., 2021; Xu et al., 2024; Sun et al., 2023; Zbontar et al., 2021; Bardes et al., 2022; Caron et al., 2020; Dosovitskiy et al., 2021; Steiner et al., 2022; Touvron et al., 2022; Liu et al., 2022; He et al., 2016). This classification is also the objective dummy used in every partial correlation. Section 5.5 and Figure 3 split embedding prediction (I-JEPA) out of the masking row, since it carries the opposite fingerprint.

encoders at 518 preserves the significant diagonal associations ($+.52$–$+.63$) and the positive BInD metric association ($+.58$). TokenEnt remains negative ($-.51$) but no longer survives the correction. DINOv2's arrangement reading changes from .672 to .498, emphasizing the need to state resolution rather than treating it as immaterial.

Capability controls and model dependence. ImageNet kNN top-1 uses ten gallery images and five queries per class; scores range from .065 to .849. A larger gallery of 50 training images per class with ten validation queries gives nearly the same ranking (Spearman .997). Under that control, BInD and TokenEnt remain significant at $+.58$ and $-.58$. Aggregating size variants into 17 training recipes gives partial correlations of $+.66$ and $-.83$. This reduces the influence of repeated size variants but does not eliminate dependence between related training recipes.

Controls and correction. Both variables are regressed on objective indicators and, where specified, the kNN score; the residuals are correlated with $\mathrm{df} = n - 2 - (\text{number of controls})$. Test families use Benjamini–Hochberg correction at $q = .05$. The indicators follow Table 8; the descriptive I-JEPA split in Figure 3 does not change that control taxonomy. Reassigning EVA02 from masking to language alignment leaves BInD at $+.54$ and TokenEnt at $-.52$, both significant.

Compression. ToMe pairs similar tokens and aggregates them. Its official proportional attention weights token contributions by their size; the ablation removes that weighting. We refer to this existing component as token-mass accounting, not as a new method. Pruning defaults to CLS-attention saliency, with feature-norm and random variants explicitly identified. The sweep covers DINOv2, DINOv3 (Siméoni et al., 2026), CLIP, DeiT3, and MAE.

Distillation. The supervised ViT-S student is distilled for 5,000 steps with four axes measured every 500 steps, including initialization. One contrast changes the teacher; the other changes the alignment objective. Appendix I reports the seed ranges.

Vision-language models. The systems are LLaVA-1.5-7B (Liu et al., 2023; 2024) (CLIP ViT-L/14-336 with Llama-2) and LLaVA-Interleave-0.5B (Li et al., 2025a) (SigLIP so400m (Zhai et al., 2023) with Qwen1.5 (Bai et al., 2023)). BInD reads the mean of the second-to-last vision hidden states after removing CLS. The language model consumes the tokens individually, so this mean is an evaluation readout rather than an identical system interface. The frozen stimuli are upsampled to deployment resolution. Two-alternative caption-likelihood evaluation uses the attribute-swap subset of SugarCrepe (Hsieh et al., 2023), with 666 COCO (Lin et al., 2014) items. Retrieval uses the same vision tower as a dual encoder with its text tower on those items. These tasks are distinct from open-ended generated-answer VQA.

E DESIGN CHOICES AND MEASUREMENT SCOPE

Near-hue design. Distinct colors put 19 of 32 encoders above .95 on the color axis. Similar hues restore score variation; distinct-color variants remain available as ceiling checks. Their ranking agrees with the near-hue ranking at Spearman .79. This is evidence of some stability across difficulty settings, not invariance to hue span.

Estimator comparison. With objective and capability controlled, the broken response correlates with color accuracy at $+.54$, the preserved response at $+.22$, the ratio at $+.50$, and the difference at $+.56$. The small difference between $.54$ and $.56$ does not establish predictive superiority. The difference is chosen to express the preserved/broken contrast without a small ratio denominator.

Pixel-distance control. On the margin stimuli, broken-edit pixel cosine distances range from .028 to .039. The preserved distance is within .005 on the binding axes; their difference is within $\pm.005$ and slightly negative on color and load. Thus this particular pixel baseline does not explain a positive BInD score. Equal pixel distances do not imply equal nuisance responses in every encoder. For size, the preserved edit has a smaller pixel change; we use size accuracy as a control and make no binding-specific claim from its margin.

Hue-discrimination control. On a one-object task with the same hue pairs and position nuisance, encoders score .82–1.00 while the pixel baseline is at chance. The association with the color axis is $+.24$ and nonsignificant. These results make simple inability to distinguish the hues an insufficient explanation of the full color pattern; they do not exclude all feature-level alternatives.

Synthetic versus photo sensitivity. At moderate reduction, the photo suite reports 94–100% preservation for three merge variants, while the synthetic suite reports 62%, 77%, and 117%. The synthetic protocol distinguishes those settings, but photographs do not confirm comparable loss at that reduction. We therefore interpret these synthetic differences as diagnostic warnings rather than established real-image degradation.

Readout dependence is not information loss. An invertible change of representation preserves recoverable information, whereas a non-orthogonal linear transform can change cosine distances and their ordering. Consequently, BInD is not invariant to all information-preserving reparameterizations. Likewise, raw pixels contain the rendered scene even when the pixel-cosine baseline fails the curated forced-choice test. A low BInD score therefore concerns accessibility under the prescribed readout, not the absence of all binding information.

A designed-representation check. Our margin–accuracy comparisons share an intervention principle, so we add a check whose answer is fixed by construction. From the item specifications alone we build two synthetic encoders, one embedding the multiset of attributes with the pairing erased and one embedding the shape–color conjunctions, and mix them with weight α. The attribute bag reads .53 and .50 on the color and load axes with a margin of .0001; the conjunction code reads 1.0 with a margin of .90; and both readouts rise monotonically in α, the margin through the full range, from .0005 at α=.02, while the accuracy saturates by α=.25. The score tracks exactly the designed difference, pairing present against inventory alone; a decoder-based comparison on learned encoders remains open.

F THE METRICS WE COMPARED AGAINST

All eleven are measured on ImageNet validation images, five per class, through each model's own preprocessing at 224.

Spectral. PR is the participation ratio of the covariance eigenvalues, the squared sum over the sum of squares. Effective rank exponentiates the entropy of the same spectrum. RankMe computes it from singular values. Alpha-ReQ is the power-law decay exponent of the spectrum.

Distributional. Isotropy averages the absolute cosine similarity over random pairs and subtracts it from one. Uniformity is the log of the mean exponentiated pairwise distance on the sphere. TwoNN estimates intrinsic dimension from the ratio of first and second neighbour distances.

Stability, diversity, and sensitivity. GS splits the feature dimensions into random halves and correlates the two image-similarity structures they produce, averaged over 30 splits on 1,000 images. TokenEnt is the entropy of the patch-token similarity matrix within an image, over 200 images. LiDAR measures how well augmented views of one image are grouped, and takes the effective rank of that discriminant. JER is the effective rank of the input-output Jacobian, over 32 directions and 100 images.

JER needs an exact Jacobian. Finite-difference and exact Jacobian-vector-product estimates differ substantially in our check: on Barlow Twins they give 3.8 and 25.8, respectively, compared with 29.3 in the original report. Crossing activation boundaries is a possible source of this discrepancy. All JER results here use exact Jacobian-vector products.

Raw correlations. Before any control, against the color binding of the 32 encoders, PR reads $+.58$, the BInD margin $+.54$, and JER $+.10$, and the margin tracks the kNN capability itself at $+.65$.

The suppression case. TokenEnt correlates with color accuracy at $-.11$ before controls and at $-.65$ after controlling objective. This change illustrates statistical suppression in this model set. It does not identify a causal pathway from objective or token diversity to binding.

G WITHIN-RECIPE METRIC INTERVENTIONS

Supervised recipe and seed variation. A pretrained ViT-S continues supervised cross-entropy for 200 steps. Classification is top-1 on a fixed 1,000-image validation set. Each counted arm uses four seeds and adds one differentiable metric term to the otherwise matched training protocol. The control binding scores span .45–.53; no counted arm's mean differs from the control mean by more than that span's width (Table 9). Within-arm seed differences reach nine points.

Arm	Direction	Mean ± SD	Seed values
Control	—	$.486 \pm .033$	.530, .490, .470, .453
JER	up	$.518 \pm .032$	.567, .503, .507, .497
JER	down	$.508 \pm .044$	.540, .473, .550, .467
PR	up	$.480 \pm .026$	.510, .457, .493, .460
PR	down	$.542 \pm .019$	.570, .537, .527, .537
Effective rank	up	$.483 \pm .043$	.547, .453, .467, .463
Effective rank	down	$.565 \pm .033$	.543, .543, .613, .560
TokenEnt	up	$.530 \pm .038$	.490, .510, .543, .577
TokenEnt	down	$.506 \pm .036$	.537, .467, .483, .537
Alpha-ReQ	up	$.474 \pm .043$	.533, .440, .447, .477
Alpha-ReQ	down	$.475 \pm .035$	.513, .430, .470, .487
Uniformity	up	$.506 \pm .028$	.537, .493, .473, .520

Table 9: **Binding scores per arm in the supervised recipe.** Four-way distinct-color accuracy after 200 steps; each arm lists its mean, the standard deviation across seeds, and the four seed values. Isotropy is excluded from counted comparisons for classification collapse, and uniformity is evaluated upward only.

These are descriptive ranges, not evidence that every practically important effect is absent. A cross-model predictive relationship also need not remain valid under direct optimization of the metric.

Intervention battery and positive controls. The battery has 300 items per axis, deterministically generated from the frozen specification and a fixed seed at training resolution. It uses a distinct-color four-way task rather than the frozen near-hue task, because the anchor starts at .53 rather than at ceiling. An arm targeting the margin drives four-way accuracy from .53 to chance, while a contrastive control reaches .79 in 200 steps. These controls show that the battery can register

training-related changes. They do not establish sensitivity to every binding effect or every stimulus regime.

Added near-hue check. For TokenEnt descent, the near-hue task ends at .86–.91 across four seeds, compared with .90–.95 for controls. There is no observed gain; the ranges alone do not demonstrate equivalence to the control. The margin-targeted arm that drives four-way accuracy to chance leaves near-hue performance at the control level. This dissociation limits transfer between batteries and should not be described as general readout invariance.

Metric	Increasing arm	Decreasing arm
JER	$2.0\times$	$0.07\times$
PR	$1.5\times$	$0.01\times$
Effective rank	$1.1\times$	$0.01\times$
TokenEnt	$2.1\times$	$0.07\times$
Alpha-ReQ	$2.4\times$	$0.54\times$
Uniformity	$0.16\times$ (reported)	floor

Table 10: **Reported metric changes in the supervised recipe.** These factors summarize metric movement, not uncertainty in binding effects. Uniformity has a floor and is assessed in one direction.

Magnitude and scope of metric movement. Classification ends at .70–.80 in counted arms. PR reaches 1.6, below the frozen-set range of 2–640, while classification remains .775. Alpha-ReQ stays within its observed encoder range throughout both arms; effective-rank descent crosses that range at its midpoint. The reported binding changes stay within the observed seed ranges at the evaluated checkpoints. All metrics except TokenEnt use the pooled features used for binding; TokenEnt uses the underlying patch tokens. Optimizing one spectral metric also changes others, so the intervention is not an isolated manipulation of a single representational property.

Excluded and untested arms. Isotropy descent reduces classification to .08 and is excluded from capability-preserving comparisons. Its ascent starts at .94 near a .98 ceiling and moves by about one point, too little to assess a large intervention. Uniformity is evaluated upward only. TwoNN and GS are not included in the differentiable protocol, LiDAR requires a separate augmentation pipeline, and RankMe is not separately optimized because of its close entropy-based relationship to effective rank.

Contrastive recipe. The second recipe uses a SimCLR anchor and a twenty-class nearest-neighbour capability probe. Metric changes include effective rank down to 1.08, PR down to 1.05, and arms maintaining them at 208 and 170 while controls reach 74 and 33. TokenEnt rises $2.3\times$, alpha-ReQ moves in both directions, and JER spans a 27-fold range. These metric movements include arms with different levels of capability preservation, which must be separated.

Four arms retain probe performance within the controls' .74–.88 range. Their binding means fall within the broad control binding range, .48–.79. This does not resolve modest effects. Three further arms reduce the capability probe to .23–.52 and are not counted as clean dissociations. In particular, descent toward effective rank one accompanies probe scores of .31 and .23; it is not a capability-preserving example. Two arms, TokenEnt descent and uniformity ascent, reduce the probe to .04–.15 (chance .05) and binding to .27–.31 (four-way chance .25). The first recipe supplies the capability-preserving TokenEnt-descent comparison.

Interpretation and uncertainty. The observed metric changes do not yield a consistent binding improvement in these short training regimes. They show that the tested optimizations are not sufficient for the expected improvement; they do not prove that all effects are zero or that observational metric associations are invalid. Per-seed outcomes are reported in Table 9; prespecified equivalence margins would be needed for stronger null claims.

H TOKEN REDUCTION IN FULL

Preservation is defined by Equation 3, against the same encoder without compression. It normalizes excess accuracy above chance; it is not a fraction of total representation information. The pipeline is

deterministic given the checkpoint and stimuli, which removes run-to-run randomness but not item-sampling uncertainty. For the central comparison, the raw near-hue color accuracies behind Table 2 are .959 uncompressed, .658 for merging without proportional attention, .895 with it, and .960 for CLS-saliency pruning, on the 1,000-item two-alternative task with chance at .50; Equation 3 maps these to the reported percentages.

Arm	Constraint or change	Color (%)
ToMe, no prop. attention	No matching constraint	34
No inter-object merges	Exclude different-object token pairs	33
Random constraint control	Exclude the same number of pairs at random	40
Same label only	Also exclude object–background pairs	31
Background only	Exclude pairs touching object tokens	51
Official prop. attention	Restore token-size weighting	**86**
Pruning reference	CLS-attention saliency	100

Table 11: **Constraint ladder: DINOv2 at $r = 16$.** Object labels come from majority projection of item connected components onto the patch grid. Constraints mask the matching score. The baseline omits proportional attention; the restored setting uses ToMe's official component.

Mechanism inspection. None of the 2,112 inspected merges in the DINOv2 $r = 16$ setting joins tokens of two different objects under the stated labeling procedure. Forbidding those pairs does not recover the score. Thus direct inter-object merging is not supported as the explanation in this inspection. Background-only merging gives 51% preservation, but this comparison does not uniquely identify the residual mechanism or explain the remaining loss under proportional attention.

Encoder	ToMe, no prop.	ToMe, official	CLS-saliency pruning
DINOv2	34%, 18%	86%, 24%	100%, 85%
CLIP	−4%, 19%	82%, 68%	85%, 99%
DeiT3	69%, 59%	97%, 90%	63%, 44%
MAE	24%, 8%	79%, 43%	45%, 45%
DINOv3	18%, 37%	41%, 33%	56%, 57%

Table 12: **Color preservation across five encoders.** Entries show $r = 16$, then $r = 24$. Restoring proportional attention improves nine of ten matched cells; RoPE-based DINOv3 reverses the change at $r = 24$. The size of recovery is model- and reduction-dependent.

Saliency comparison. Table 3 uses the same frozen battery for the norm and random arms. Random pruning at $r = 16$ preserves 16%, 1%, 18%, and −2% on DINOv2, CLIP, DeiT3, and MAE, respectively. Both random and norm pruning preserve 56% on DINOv3. Official ToMe exceeds CLS-saliency pruning on DeiT3 and MAE in Table 12, so pruning is not uniformly preferable to merging.

Task scores versus color preservation. Table 2 brings the main DINOv2 comparison into the body. Official versus ablated proportional attention gives kNN .7642 versus .7643, ADE mIoU .285 versus .280, and NYU RMSE .765 versus .766, while color preservation is 86% versus 34%. Pruning has kNN .768 versus the uncompressed .785, but ADE mIoU falls from .316 to .184. These results motivate complementary reporting; they do not establish that BInD is the only diagnostic capable of detecting these changes.

I Distillation in full

Setup. A supervised ViT-S student is distilled for 5,000 steps, with measurements at initialization and every 500 steps (eleven checkpoints). One contrast varies teacher color performance; the other fixes a DeiT3 teacher at .32 arrangement accuracy against the student's initial .42 and varies alignment.

Teacher contrast. The student's color accuracy begins at .94. An I-JEPA teacher scores .71 and yields students at .73 and .78 across two seeds. The color-strong DINOv2 teacher yields .77 and

.86. This comparison is consistent with teacher-specific transfer, but its small number of seeds does not isolate all teacher differences.

Alignment contrast. Each arm uses two seeds, except standardized logits, which uses four. The arrangement ranges in Table 4 are final observed seed ranges, not confidence intervals. Distance alignment shifts the student toward the teacher, including axes on which the student initially leads. Logit alignment preserves the arrangement advantage at all eleven checkpoints. Standardized logits range from .32 to .58, while relation alignment retains the measured axes and raises arrangement. These statements describe the evaluated student–teacher comparison rather than an objective-wide guarantee.

Second teacher. A ConvNeXt-B teacher has arrangement accuracy .35 against the same student's .42. After 5,000 steps, relation alignment reaches .49 on both seeds. Logit alignment preserves the advantage at every measured checkpoint. Distance alignment ends at .33 on one seed and .44 on the other, so teacher-level convergence is not consistent across seeds. Standardized-logit trajectories range from .32 to .52. The cases suggest similar tendencies, but do not establish a statistically resolved ordering of all four objectives.

J THE DENSE READOUT

The dense protocol reuses the frozen margin stimuli (300 triplets on each of four axes) and averages the intervention difference over aligned patch positions. The downstream targets are ADE20K segmentation (Zhou et al., 2019) and NYUv2 depth (Ignatov et al., 2024). The dense column of Table 13 is the color-axis dense margin, Equation 2 on the 300 color triplets, with no aggregation across axes; the pooled counterpart below is the color-axis margin of the pooled embedding on the same stimuli. Each cell is a partial Spearman correlation, computed by rank-transforming both variables and residualizing on the training-objective dummies of Table 8, over 31 encoders; BEiTv2-L is excluded because its NYUv2 probe diverges, an RMSE of 8.7 against at most 1.1 elsewhere, and depth is scored as negative RMSE. Bold marks the cells that survive a Benjamini–Hochberg correction over the family of eight tests at $q = .05$.

Target	Static geometry (PR)	BInD, dense (color)
ADE20K mIoU	$+.\mathbf{50}$	$-.03$
ADE boundary F	$-.17$	$+.40$
ADE fragmentation	$+.38$	$\mathbf{-.56}$
NYU depth	$-.41$	$\mathbf{+.64}$

Table 13: **Dense associations across 31 encoders**, with the training objective controlled; the dense column is the color-axis dense margin, and one encoder is excluded for a diverging depth probe. Bold survives correction over the family of eight. These correlations do not establish mutually independent quality components.

The eight classification targets are ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021), Pets (Parkhi et al., 2012), Flowers (Nilsback & Zisserman, 2006), Food (Bossard et al., 2014), Aircraft (Maji et al., 2013), and iNaturalist 2019 (Van Horn et al., 2018). Global size associations range from $+.45$ to $+.71$, with five surviving correction. The pooled reading of the same color-axis margin associates with depth at $-.63$ ($p = .0006$), opposite in sign to the dense $+.64$ on the same encoders, stimuli, and controls. The depth residual is $+.20$ and nonsignificant after controlling segmentation, consistent with shared variation among geometric targets rather than an established depth-specific component.

K THE FINGERPRINT ACROSS ARCHITECTURE FAMILIES

Five families the frozen set does not cover, all trained with ImageNet supervision at the resolution they ship at, scored with the frozen suite and the same code.

Reproduction is read against a fixed criterion, both binding axes above .8, arrangement at least 35 points below the lower of them, and a negative arrangement margin, and all five meet it. Absolute

family	color	load	size	arr.	arr. margin
the supervised cluster	.892–.963	.764–.945	.794–.875	.296–.396	negative
Swin-B, window attention	.930	.931	.760	.456	−.138
MLP-Mixer-B, no attention	.948	.916	.871	.483	−.116
ResMLP-24, no attention	.947	.906	.827	.332	−.144
MambaOut-B, gated conv.	.885	.863	.778	.336	−.157
MambaVision-B, state-space	.835	.905	.739	.335	−.131

Table 14: Accuracy on the frozen suite. All five show the broad supervised-recipe pattern: high color/load accuracy, lower arrangement accuracy, and a negative arrangement margin. This does not isolate the loss function from the rest of the recipe.

levels scatter around the frozen cluster in both directions, two families above it on arrangement and two below on the binding axes, which we read as variation inside the pattern rather than as family effects, since each family is represented by one model. MambaVision is a hybrid with a small number of attention blocks at the end, which we state rather than smooth over.

The fingerprint inside one training set. Objective and training data are entangled across the frozen set, since the language-aligned encoders also see web-scale data. Thirteen encoders share ImageNet-1k as their only training images and span every objective in the set except language alignment, which is absent from this subset. Inside the subset the fingerprint reproduces. Supervised, masking, self-distillation, and clustering keep arrangement as their lowest binding axis, embedding prediction stays first on arrangement and last on color among the families, and each family with more than one member correlates with its full-set shape at .92 to 1.00. The within-profile peak of embedding prediction at arrangement belongs to the model trained beyond ImageNet-1k, so inside the subset the family leads on arrangement without peaking there. For language alignment the entanglement stands, and the fingerprint of that family is a claim about the objective and its data together. The objective label names the whole training recipe, augmentations included; what the fingerprint separates is recipes, and no claim here splits the loss term from the rest of its recipe.

L VISION-LANGUAGE MODELS IN FULL

Configuration	Color	Load	Size	Arrangement
Merge $r = 8$	66%	69%	97%	112%
Prune $r = 8$	94%	102%	109%	117%
Merge $r = 18$	62%	52%	101%	168%
Prune $r = 18$	117%	85%	123%	162%
Merge $r = 24$	5%	13%	16%	73%
Prune $r = 24$	70%	45%	69%	152%

Table 15: **CLIP ViT-L/14-336 inside LLaVA on the synthetic battery.** Values are reference-normalized preservation, not raw accuracy. At $r = 24$, merging also reduces size to 16%, precluding a binding-specific attribution of the color decline.

Raw scores, chance levels, and conventions. Each percentage in Table 5 and Table 15 is the relative change of the margin above chance; the synthetic, caption-likelihood, and retrieval readouts below are two-alternative tasks with a chance level of .50. Table 16 gives the raw accuracies behind the extreme-compression comparison; the photographic paired-composite row of Table 5 uses a separate sample and a McNemar comparison, described below.

Relative changes are task-specific. Ratios of the reported synthetic-color and caption-likelihood relative decreases are 17× at $r = 8$, 9.5× at $r = 18$, and 3.3× at $r = 24$. These descriptive ratios compare different tasks and scales; they are not calibrated measures of how much common information a benchmark misses. For pruning, color preservation is at least 94% at the first two rates. The moderate merge settings retain the size control, whereas the extreme setting does not.

Stimulus and readout comparison. Synthetic intervention preservation is 62% at $r = 18$ and 5% at $r = 24$. The real-photo intervention gives 96% and 57%, while the photographic behavioral

Readout	n	Uncompressed	Merge	Official ToMe	Prune
Synthetic color, tower	1000	.848	.516	.517	.745
Caption likelihood	666	.815	.725	.752	.767
Retrieval, attribute swap	666	.632	.616	.632	.625

Table 16: **Raw accuracies at extreme compression ($r = 24$) for the LLaVA-1.5-7B tower.** Two-alternative readouts with chance at .50. Merge omits proportional attention; official ToMe includes it. The synthetic reading falls to the chance floor while the behavioral readouts remain above it.

readout gives 108% and 87%, neither significant. Both stimulus material and readout affect the comparison. These values do not isolate a readout-only explanation, and the moderate photo result does not confirm the synthetic loss.

Photographs across axes. At $r = 24$, merging gives real-photo preservation of 57% for color, 68% for load, 71% for size, and 47% for arrangement; the reported changes are significant at $n = 400$. Pruning gives 86–97% across those axes. The photo row in Table 5 uses a separate set of paired composites and a McNemar comparison, not this 400-item suite. Thus its 60% decrease and this suite's 43% color decrease refer to different samples/protocols. Photographic arrangement also has a pixel-cosine shortcut (Appendix M).

Pooling alternative. Merging changes token correspondence and the weights represented by each surviving token. The unweighted mean used for BInD is not the individual-token representation consumed by the language model. Pruning's stronger score under the same mean does not rule out a merge-specific pooling artifact. We compare the unweighted mean with the token-mass-weighted mean,

$$\bar{z}_{\text{uniform}} = \frac{1}{m}\sum_{j=1}^{m} z_j, \qquad \bar{z}_{\text{mass}} = \frac{\sum_{j=1}^{m} s_j z_j}{\sum_{j=1}^{m} s_j},$$

where s_j is the number of original patches represented by token j. The two poolings coincide without reduction, and the mass-weighted mean does not recover the color reading at any reduction, so the decline of the mean is not an artifact of ignoring token mass. A token-aware check through the language model remains open.

Retrieval comparison. For the ViT-B/16 retrieval system at $r = 16$, attribute-swap performance declines 1.6–3.0 times as much as its object-replacement control under the reported relative-change convention. The merging comparison is significant at $p = .009$. This supports a differential effect on those annotated retrieval tasks; it does not calibrate a universal damage scale.

Second architecture. For SigLIP so400m with Qwen1.5, reduction leaves 105 of 729 tokens at the extreme setting. The reported relative-change ratios are $6.2\times$ and $4.0\times$ for merging at $r = 12$ and $r = 24$, and $11.4\times$ and $3.1\times$ for pruning. They retain the task-scale caveat above. The uncompressed system scores .805 compared with .815 for the 7B system. Similar baseline performance does not by itself isolate architecture, language-model compensation, or reduction depth as the explanation for different compressed responses.

M THE REAL-PHOTO ANCHOR

Construction. Objects with verified masks are cut from photographic sources and composited, and the color factor is a hue rotation applied only inside the object mask. The swap rules and the hygiene battery are the synthetic ones; the pixel-only examinee sits at chance on color, load, and size, and solves arrangement at .99, since photographic texture makes position matchable from the pixels alone. The arrangement transfer below is therefore reported and not leaned on, and the supported rank-transfer claims in Section 3.5 concern size and load.

Rank transfer. With the objective controlled, the ranking of the encoders transfers from synthetic to photographs at $+.77$ on size, $+.66$ on load, and $+.56$ on arrangement, against an agreement of .78 to .91 between photographic sources themselves. Color transfers at $+.30$ and is not significant, while the photographic sources agree with each other at .78 on that axis.

Hybrid stimulus checks. We built two hybrids, one with photographic silhouettes and flat palette colors and one with synthetic shapes and photographic color texture. Both hybrids show rank transfer of $+.67$–$+.68$. This supports sensitivity to stimulus material, with substantial transfer in both hybrid settings. It does not prove that synthetic and photographic color scores measure an identical latent construct, or identify a unique cause of the original gap.

A method note. A pilot at n=12 showed the opposite pattern, which disappeared at n=32. Transfer verdicts in this appendix are only reported at $n \geq 30$.

What ships. Recipes rather than pixels, at 916KB, with four tasks of 400 items, the cutout provenance as dataset, image id, and box, the alpha masks, and the code that rebuilds and composites. Rebuilt items score within 0.02 of the originals on every task and encoder we checked.

N WHAT MMVP RECORDED, READ THROUGH THE PROFILE

Design. Pair discrimination is $1 - \cos$ between the two embeddings of an MMVP-VLM pair, over 135 pairs in nine categories. It needs no text tower, so all 32 encoders can be scored, where the original report could only score the CLIP family.

Two steps that the measurement needs. Raw discrimination is contaminated by scale, since MAE-H reads 0.002 and looks completely blind until it is divided by a random-pair baseline of the same category, where it reads 0.50. Every number below is the normalized quantity. And because the pairs were selected with CLIP and DINOv2, the main analysis excludes both families at n=25. The correlations strengthen after exclusion. This reduces one specific dependence on the pair-selection models, but does not eliminate all selection effects.

axis → category	r	axis → category	r
size → Color	$+.83$	arrangement → Quantity	$+.67$
arrangement → Camera Perspective	$+.82$	size → Presence	$+.65$
arrangement → Structural	$+.77$	arrangement → Spatial	$+.65$
arrangement → Presence	$+.69$	size → Camera Perspective	$+.63$
arrangement → Color	$+.68$	size → Structural	$+.62$

Table 17: Everything that survives correction at n=25 with the objective and the capability controlled; the ten were fixed on one half of the photograph pairs, and nine held on the other. All ten come from arrangement or size, and neither color nor load meets the criterion; the color-axis correlation with the Color category is $-.07$. Nonsignificance does not establish that binding is irrelevant to the category.

Split-half confirmation. The ten were fixed as hypotheses on one half of the pairs and retested on the other half with a signed one-sided test. Nine held, and the one that did not is arrangement to Presence.

Embedding-level discrimination. Normalized discrimination is .209 for EVA02-B, .217 for EVA02-L, .227 for EVA-CLIP, .249 for CLIP-L, and .267 for CLIP-B16. Higher readings include I-JEPA at .532, Barlow Twins at .560, and VICReg at .560. These are embedding-distance comparisons on the selected pairs, not direct measures of multimodal-system correctness or unique deficits in binding.

Limits. Discrimination is an embedding-level measurement, and the chain from there to the behaviour of a multimodal system belongs to the original report. The sample is 25 encoders and 15 pairs per category. And the predictive power of the size axis may share a geometric component with discrimination itself.

O A REPLICATION ATTEMPT ON THE CLIP FAMILY

Design. We examine fifteen CLIP variants from four data sources and three scales. The target is a sample-weighted SugarCrepe (Hsieh et al., 2023) composite of 7,511 trials. JER uses its reported probe and formula. The JER and BInD rows control architecture, coded by variant, and ImageNet zero-shot accuracy. The ImageNet row controls architecture, coded by variant. Significance is a two-sided t test on the partial correlation at $p < .05$, with n per row set by measurement availability (15 for all rows shown). The original variant set is not public, limiting exact replication.

Predictor	Raw	Controlled
JER, natural images	$-.05$	$+.31$, n.s.
ImageNet zero-shot	$+.77^{*}$	$+.67^{*}$
BInD, color accuracy	$-.35$	$-.38$, n.s.

Table 18: **Within-family composite associations.** The JER and BInD rows are controlled for architecture and ImageNet zero-shot, and the ImageNet row for architecture. Neither JER nor BInD color accuracy is significant in this sample, and ImageNet zero-shot is positively associated with the composite.

Interpretation. The reported JER effect does not reproduce on these variants. This is a failed replication in our setting, not a general refutation of JER. ImageNet's positive association shows a substantial capability component in the composite; it does not show that the composite contains no binding-related variation. For isolated binding-specific subtypes with their control tasks held fixed, neither JER nor BInD meets the reported criterion. Limited within-family variation and the small model sample constrain this inference. They do not make within-family validation inherently invalid or prove that no metric can predict those subtypes.

Implications for our readout. Within this family, the margin correlates with capability at $-.83$. Four attempted normalizations do not resolve that association; the preserved-edit response tracks capability. Color accuracy has an association of $-.16$, which is nonsignificant. This motivates using accuracy for the reported cross-encoder comparisons, but does not prove that accuracy is universally independent of capability.

P CASE STUDY, DINOV2 AGAINST DINOV3

Why this pair. It is the comparison a practitioner faces, two uncompressed models, and it is also where the resolution rule earns its place.

axis	v2 at 224	v2 at 518, its own	v3 at 224	v3 at 256, its own
color	.959	.924	.818	.844
load	.960	.845	.711	.838
size	.980	.993	.976	.965
arrangement	.672	.498	.802	.725

Table 19: A shared resolution distorts in both directions. At 224 it reads a 25 point win for v2 on load, which becomes a 0.7 point difference once each model is scored where it ships.

The fair comparison is a trade, not a ranking. At their own resolutions v2 leads on color by 8 points and slightly on size, load is a tie inside the resolution error of $\pm.022$, a resolution-sensitivity bound distinct from the item-sampling error of Appendix C, and v3 wins arrangement by 23 points. Rescoring v2 at 518 also lowers raw arrangement accuracy to .498; the four-choice chance level is .25.

Post-hoc relation to design intent. DINOv3 uses an anchoring term intended to preserve patch-space structure (Siméoni et al., 2026). Its higher native-resolution arrangement score is compatible with that motivation, but the present comparison is pooled and does not directly test patch-level structure. The resolution dependence of DINOv2 may have several explanations; this case study does not identify an artifact-token mechanism.

Scope. This is an illustrative comparison of deployed configurations, not a preregistered prediction or an isolated architectural intervention. Patch sizes differ (14 versus 16), as do native resolutions. The current DINOv3 implementation does not support the corresponding dense readout, so no DINOv3 dense-validation result is claimed here.

Q WHAT AN ITEM LOOKS LIKE

Each item is a forced choice. The query sits on the left, the candidates follow, and the candidates hold the same inventory as the query so that only the pairing tells them apart. The correct candidate is marked here and is unmarked in the suite.

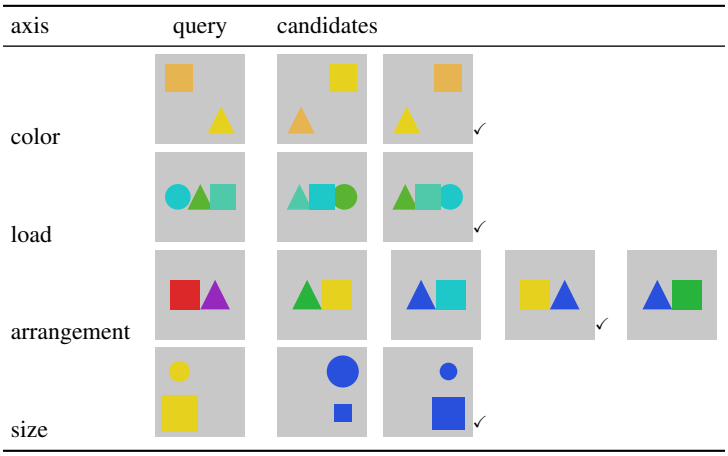

Figure 4: **One item from each axis of the released suite.** Every axis keeps the candidates' attribute inventory fixed and varies only the queried pairing, while the remaining attributes are redrawn between query and candidates so that appearance matching cannot identify the answer. Color and load hold the colors and shapes and swap which color sits on which shape; load adds a third object. Arrangement keeps the objects and the positions, varying which object sits where over four candidates. Size fixes the shapes and both sizes; the candidates differ in which object is larger, and it serves as the control axis. The check marks the correct candidate.

The colors of the color and load axes are near-hues rather than distinct ones, which is what keeps the axis from saturating (Appendix E), and it is why the two candidates of a color item look alike at a glance.

R USE OF LARGE LANGUAGE MODELS

A large language model assisted implementation and refactoring of measurement and figure code, language revision, and literature search. Manuscript revision also used it to reorganize sections and tables, improve readability, and check whether the wording of claims matched the reported evidence. This revision did not generate experimental measurements or add results for unperformed validation studies. The authors are responsible for inspecting and verifying the implementation, numerical results, references, interpretation, and final manuscript.

S ETHICS STATEMENT

The suite is generated from specifications rather than collected, so no human subjects and no personal data are involved on the synthetic side. The real-photo anchor composites objects cut from existing datasets, and we ship provenance and masks rather than pixels, so it inherits the licences of its sources and redistributes none of their images. The instrument is diagnostic and we see no application of it that creates a new capability for harm. Its main foreseeable misuse is the one we warn about in the paper, which is treating a single score as a verdict on an encoder, and the failure mode of our own instrument is the same, which is why Section 3.4 and Appendix B state what a reading does and does not license.

T REPRODUCIBILITY STATEMENT

Section 3.1 defines the measure and Section 3.3 specifies the suites. Appendix C describes generation, curation, and freezing; Appendix D describes scoring and experimental setup. Controls and correction appear in Section 4.2 and Appendix D; the within-recipe interventions appear in Appendix G. The metric-intervention runs use a deterministically regenerated training-resolution battery rather than the frozen files, with the additional near-hue evaluation explicitly distinguished. Real-photo, MMVP, and downstream evaluations use the datasets and protocols identified in their respective sections. The release comprises item specifications, generation/scoring code, checksums, pinned versions, and numerical tables.