# Sparse Groundings: A Benchmark for Auditing Visual Representation Claims

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Modern multimodal systems rely on internal visual representations, but claims about what those representations contain are typically supported by probes or interventions read from the patches covering an object. These readouts can conflate object evidence with signals that are easier to miss: information broadcast across the image, background cues, patch coordinates, or the evaluator's own pooling rule. We introduce Sparse Groundings, an audit benchmark for visual representation claims built around three controlled readouts at the same patch count. The standard object readout is compared with a blank same-footprint readout that preserves selected patch coordinates while removing image content, and a matched-background readout that preserves image context and patch count while removing the object. Each result is recorded as a claim card summarising the tested property, controls, method, and verdict, and aggregated into an evidence profile rather than a single groundedness score. We instantiate the benchmark on controlled tests for identity, colour, position, scale, spatial relations, counting, and distractor robustness across six vision encoders and two sparse-autoencoder bases. The controls change conclusions in three places. In one representative case, a SigLIP 2 probe predicts which cell of a $7 \times 7$ grid contains the object with $0.991$ balanced accuracy from object features, while a uniform blank image read from the same patch locations reaches $0.998$ — a pattern that holds at $\geq 0.978$ across all eight representations. Identity and attribute content is recoverable from non-object patches as well as from object patches on every encoder, and structure-family claims partition encoders on a different axis from position-family claims. A natural-image bridge readout on LVIS object boxes recovers the same encoder partition — five of six encoders show object-pool features above the strongest control, with MAE the same matched-background-dominated outlier — so the regimes are not artefacts of the grey-canvas protocol. Task-matched evidence profiles align with ADE20K segmentation and NYUv2 depth, do not predict iNat fine-grained classification, are fragile on FSC-147 counting, and offer evidence for RefCOCO+ grounding under the current pooled-feature protocol. Sparse Groundings therefore audits not whether a model is grounded in general, but which visual claim survives which controls, by which method, and for which downstream use.

## 1 Introduction

Modern vision encoders publish localisation, dense prediction, and transfer as core capabilities. DINOv2 validates frozen features across image-level and pixel-level benchmarks [31]; SigLIP 2 reports improved localisation, dense features, and transfer when used as a VLM vision encoder [49]; DINOv3 [40] introduces Gram anchoring to maintain dense feature quality across long self-supervised training; MAE [17] achieves strong patch-level reconstruction; and CLIP [36] is a default

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

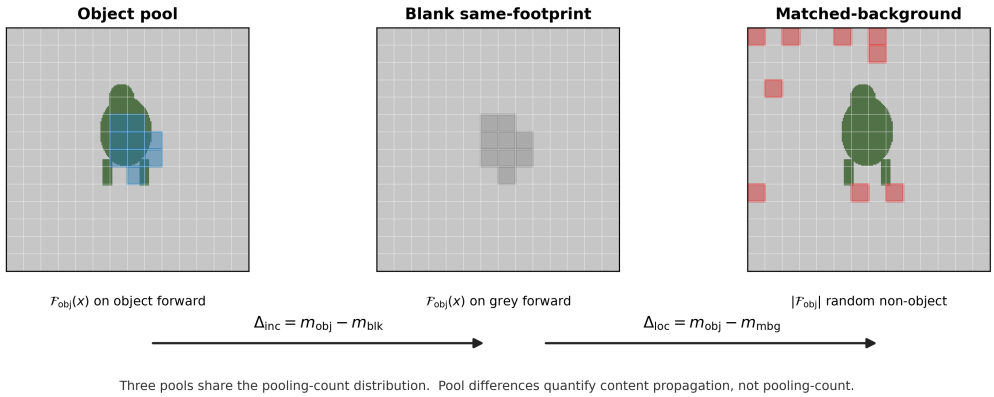


Figure 1: **The three-pool protocol.** One stimulus image, three pooling masks of equal size $|\mathcal{F}_{\text{obj}}|$. *Left:* the object pool $\mathcal{F}_{\text{obj}}$ marks patches covered by the object footprint on the original image (blue overlay). *Middle:* the blank same-footprint pool reuses the same patch indices on a uniform-grey forward (no object content). *Right:* the matched-background pool selects $|\mathcal{F}_{\text{obj}}|$ random non-object patches on the same image (red overlay), keeping context but removing the object. $\Delta_{\text{inc}} = m_{\text{obj}} - m_{\text{blk}}$ measures the incremental signal over the patch-coordinate scaffold; $\Delta_{\text{loc}} = m_{\text{obj}} - m_{\text{mbg}}$ measures object-pixel locality against background patches in the same image. All three pools share pooling count, so differences reflect content propagation rather than pool-size effects.

visual backbone for grounding stacks. These evaluations establish useful features, but they do not by themselves audit the corresponding representation claim. When the claim is tested by a probe on patch features pooled at object-relevant locations, the pooling operation is itself position-dependent: which patches enter the pooled vector is a function of where the object is. A probe trained on such features mixes two contributions — what the encoder represents about the property of interest and what the pooling operation reveals about which patches it selected.

A direct control shows this dependency. A probe that predicts which of 49 object-placement cells the object occupies, trained on a uniform-grey image pooled at the object's footprint indices, reaches at least $0.978$ balanced accuracy across all eight representations. Scale decoding clusters within three percentage points of a patch-count baseline that ignores feature content (§4). Object-pool probing therefore audits encoder content and pooling protocol jointly, not encoder content alone.

Sparse Groundings audits each visual representation claim through three pooling readouts at the same patch count: the object footprint, the same patch indices on a uniform-grey image, and a deterministic random selection of non-object patches in the same image. Two scalar deltas (object-incremental and object-locality) and a four-way regime label summarise the joint pattern. Each result is stored as a claim card. The protocol applies to any spatial token map that preserves image-region correspondence and extends the plausible-replacement discipline of Parodi et al. [34] to the per-claim, per-metric level.

**Contributions.**

1. **A three-pool audit protocol with nuisance-matched controls.** Each visual claim is read out through the object pool, a blank same-footprint pool that preserves patch coordinates while removing image content, and a matched-background pool that preserves patch count and image context while removing object pixels.
2. **A controlled module suite covering What, Where, and Structure.** Seven modules (W1 instance, W3 colour; R1 position, R2 scale; H1 relations, H2 counting, H3 distractor robustness) instantiate the protocol on synthetic stimuli with module-specific feature-free nuisance baselines, plus a natural-image bridge readout on LVIS object boxes that recovers the same encoder partition.
3. **Per-metric evidence profiles and claim cards.** Two scalar deltas ($\Delta_{\text{inc}}$, $\Delta_{\text{loc}}$) and a four-way regime label per metric describe each (encoder, module) record. Each record is exported as a claim card with controls, scores, regime, and reproduction metadata.

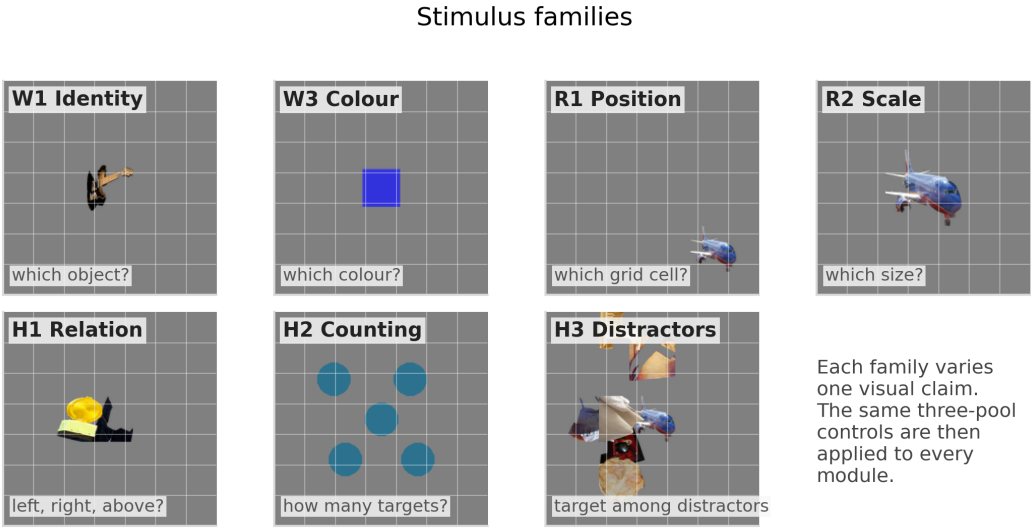


Figure 2: **Stimulus families in Sparse Groundings.** Each module varies one visual claim on a controlled grey canvas: object identity and colour (*What*); position and scale (*Where*); and relations, counting, and distractor robustness (*Structure*). The same object, blank same-footprint, and matched-background control pools are then applied to every module. W1, W3, R1, R2, H1, and H3 panels are drawn from cached generated stimuli; the H2 counting panel is shown as a schematic counting example.

4. **Task-conditioned downstream validation.** A locked selector registry maps each claim family to a claim-conditioned ranking. Spearman rank-correlation gaps against the naive object-pool baseline are positive and panel-stable on ADE20K segmentation, panel-fragile on NYUv2 depth, negative on iNat fine-grained classification, and near zero on FSC-147 counting (§6). Each outcome diagnoses whether the claim-conditioned evidence profile is aligned with the task family.
5. **A released artifact and method-extension axis.** The artifact (§7) packages the library, CLI, stimulus generators, evaluation cards, locked selector registry, and 73 schema-validated claim cards. It supports plug-in probing methods (M1–M3 naive, controlled, and top-$k$ subspace probes are reported across six raw encoders) and audits over transformed token bases; small-scale SAE feature-ablation receipts under plausible replacement [34] are included as examples (§5).

**Benchmark object.** *Inputs:* an audited spatial representation (vision encoder or token-aligned basis) and a module manifest with object footprints. *Outputs:* per-metric object, blank same-footprint, and matched-background scores; $\Delta_{\text{inc}}$ and $\Delta_{\text{loc}}$; module-specific nuisance baselines; a regime label (object-local, borderline object-local, image-wide, patch-position dominated, no readout); and a claim card. *Baselines:* the naive object-pool linear probe, the blank same-footprint and matched-background pooling controls, and the module-specific feature-free nuisance predictor. *Extensibility:* a new encoder requires a patch-token adapter; a new module requires footprints, a manifest, and a nuisance baseline.

## 2 Related work

**Output-level visual evaluation.** VLMs systematically fail basic visual perception [13, 47]; many VLM benchmarks do not require visual content at all [9]; vision-encoder choice dominates MLLM performance [46]; contrastive VLMs behave like bags-of-words on attribution and order [52]. These results identify behaviour gaps, but a behaviour gap on a downstream task does not localise which internal visual state failed or which control would have caught it.

**Controlled probing and interventions.** Control tasks for language probing [18], saliency falsifiability [1], and plausible-replacement interventions [34] discipline probe and intervention claims by

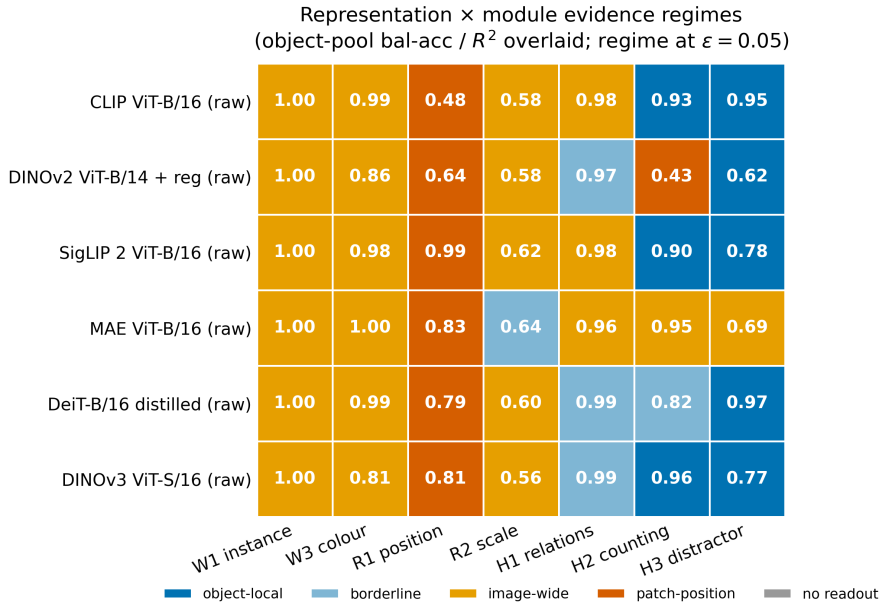


Figure 3: **Encoder × module regime heatmap.** Six raw base encoders over the seven shipped modules: W1 instance, W3 colour, R1 position (49-way bal-acc), R2 scale-v2, H1 relations, H2 counting v2, and H3 distractor robustness. Cell colour is the audit verdict at $\varepsilon = 0.05$, with a borderline shade for object-local entries within $\delta = 0.025$ of threshold; the overlaid number is the naive object-pool score, shown only to make visible that high object-pool numbers can land in different regimes. SAE-basis rows are reported in Table 1 and Appendix A.5. H2 regime labels use $\Delta_{\text{inc}} = \text{object} - \text{blank}$; under the dual-nuisance rule that includes footprint area, every encoder on H2 is below threshold (Appendix Table 9).

demanding a control. Vision-encoder representation analyses [11, 33, 37, 39] and SAE-based vision interpretability [16, 27, 32, 43] have begun to bring this discipline to vision encoders. The present paper extends it to the per-claim level: each visual representation claim is paired with two pooling controls that remove the same content the claim is supposed to be carrying.

**Claim-conditioned audit outputs.** In the spirit of mechanistic-interpretability benchmarking [29], the suite treats each (encoder, module) record as a claim that survives or fails named controls, and validates downstream against five frozen-encoder tasks. Sparse Groundings differs from VLM-behavioural benchmarks in three ways: it audits the vision substrate before the language interface; the protocol applies to any visual representation pooled over image regions; and per-claim regime labels identify the failure mode (object-local, image-wide, or patch-position dominated) rather than only its existence. Extended discussion is in Appendix A.14.

## 3 Sparse Groundings audit protocol

The unit of audit is an *audited representation*: a spatial visual state that aligns to image regions and admits pooling over object footprints. When we write "the encoder represents $X$," this is shorthand for "$X$ is recoverable from the audited representation by a controlled readout whose nuisance baselines fall below threshold"; the statement is a property of the representation under our pools, not a claim about whether a downstream system uses $X$ internally. Downstream-task validation (§6) is the only point at which representation- level findings are linked to system-level behaviour.

We instantiate the protocol on two kinds of audited representation, reported separately throughout. A *base representation* is the patch-feature output of a raw vision encoder; we use "encoder" for the backbone. An *SAE basis* is a sparse-autoencoder learned over the activations of one specific base representation; SAE results are properties of the basis, not the encoder, and are reported in a separate row labelled "SAE" next to the matching base representation.

The benchmark is encoder-agnostic: any audited representation that admits patch-level pooling over image regions can be plugged in. We treat the six-base panel below as the benchmark validation panel because it has complete coverage on R1 (the position module that surfaces the strongest control effect) and on every v2 module; the twelve-representation extended panel in Appendix A.6 is a v2-only extension and is labelled accordingly. The validation panel spans contrastive (CLIP ViT-B/16 [36], SigLIP 2 ViT-B/16 [49]), self-supervised (DINOv2 ViT-B/14 with registers [11, 31], DINOv3 ViT-S/16 [40]), masked-reconstruction (MAE ViT-B/16 [17]), and supervised-distilled (DeiT-B/16 [48]) pretraining objectives, plus two SAE bases trained over CLIP and DINOv2-reg (saev CLIP-B/16 24K and saev DINOv2-reg ViT-B/14 24K [43]). All ViT-class encoders are evaluated at block 10 with CLS and register tokens stripped before pooling; the extended panel and layer choices are in Appendix A.7.

A module encodes one claim type (identity, position, relations, and so on) through a synthetic stimulus design that varies the property along a controlled axis. Synthetic stimuli are used because natural images confound photographer center-bias with background co-occurrence. Each manifest carries footprint metadata so the three pools below can be computed deterministically.

The three pools (object $\mathcal{F}_{\text{obj}}$, blank same-footprint $\mathcal{F}_{\text{blk}}$, matched-background $\mathcal{F}_{\text{mbg}}$) are formally defined later in this section. Some modules carry an additional *module-specific nuisance baseline* — patch count for scale, independent object centroids for relations, footprint area for counting — that records what a feature-free predictor recovers from the stimulus geometry alone. Per metric $m$ on a (encoder, pool) we report $m_{\text{obj}}$, $m_{\text{blk}}$, $m_{\text{mbg}}$, the module-specific nuisance baseline if any, and two deltas:

$$\Delta_{\text{inc}}(m) = m_{\text{obj}} - m_{\text{blk}}, \tag{1}$$

$$\Delta_{\text{loc}}(m) = m_{\text{obj}} - m_{\text{mbg}}. \tag{2}$$

The two deltas are reported separately rather than combined via $\max(m_{\text{blk}}, m_{\text{mbg}})$ because that combination obscures the qualitative distinction between "signal lives in image-wide attention propagation" ($\Delta_{\text{loc}}$ small, $\Delta_{\text{inc}}$ large) and "signal is local to object pixels" (both deltas positive). A per-metric regime label summarises the joint pattern with a tolerance $\varepsilon = 0.05$ (the seed-stability scale measured under robustness checks):

- **Object-local**: $\Delta_{\text{inc}} \geq \varepsilon$ and $\Delta_{\text{loc}} \geq \varepsilon + \delta$.
- **Borderline object-local**: $\Delta_{\text{inc}} \geq \varepsilon$ and $\varepsilon \leq \Delta_{\text{loc}} < \varepsilon + \delta$. We use $\delta = 0.025 = \varepsilon/2$ as one empirical stability margin: per-cell five-seed shuffled-label probe std on R1 49-way decoding is $0.003$–$0.012$ across the audited representations (Appendix Table 16), so a $\pm 0.025$ band around $\varepsilon$ comfortably contains the seed-noise floor.
- **Image-wide object**: $\Delta_{\text{inc}} \geq \varepsilon$ and $|\Delta_{\text{loc}}| < \varepsilon$.
- **Patch-position dominated**: $\Delta_{\text{inc}} < \varepsilon$ (includes both $\Delta_{\text{inc}} \approx 0$ and $\Delta_{\text{inc}} < 0$).
- **No readout**: $\max(m_{\text{obj}}, m_{\text{blk}}, m_{\text{mbg}}) < \varepsilon$. No (encoder, module) entry in the present panel falls into this category; we keep the label for completeness and for future audited representations.

The label is a property of *a metric on an encoder*, not of the encoder as a whole; different metrics on the same encoder can carry different labels. Borderline entries are flagged separately in the heatmap (Fig. 3) and in Table 1. The methods M1–M6 are defined in §5. Each (encoder, module, method) record stores metric values, deltas, the regime label, the nuisance baseline, the manifest checksum, and reproduction metadata.

Let $\mathbf{T}_x \in \mathbb{R}^{N_p \times d}$ be the patch-token output (CLS and register tokens stripped) and $\mathcal{F}(x) \subseteq \{1, \ldots, N_p\}$ the pooling index set. The three pools share the same pooling-count distribution and reduce to mean pooling of $|\mathcal{F}|$ tokens, so differences between pool scores quantify content propagation rather than pool size. Matched-background indices are sampled deterministically per stimulus, so every result is byte-reproducible from the manifest; the exact seed construction is in §7.

Each method outputs a score per (encoder, module). The artifact schema scores each method on four axes — faithfulness against controlled-probe evidence, specificity against cross-module interference, generalisation across encoders, and concreteness [29].

# 4 Modules

Seven modules across three families are evaluated end-to-end: W1 instance recognition and W3 colour (*What*); R1 position and R2 scale (*Where*); H1 spatial relations, H2 counting, and H3 distractor robustness (*Structure*). Each module renders LVIS object catalogues [15] on a uniform $224 \times 224$ grey canvas with a $7 \times 7$ grid of object centres (49 stimulus positions; the $7 \times 7$ refers to object placement, not to the encoder's patch grid, which is $14 \times 14$ for ViT-B/16 and $16 \times 16$ for ViT-B/14). Stimulus counts and generators are in the released library; per-module result tables are in Appendices A.4 and A.13. Figure 2 shows representative stimuli from the seven modules; Figure 4 visualises the position-module control result across the six-base panel; and Figure 3 summarises the cross-module regime pattern.

**Cross-module pattern.** W1 instance recognition reaches $1.000$ on the object pool and $0.995$–$1.000$ on matched-background for every encoder, with the blank canvas at chance. Identity and attribute content is recoverable from non-object patches as well as from object-footprint patches (Appendix Tables 7, 17; W3 colour follows the same pattern), so identity and colour content is broadcast image-wide rather than concentrated in object pixels. R1 position splits into three regimes (next paragraph). R2 scale-v2 (footprint-fixed) labels five encoders image-wide and one (MAE) borderline object-local at $\Delta_{\text{loc}} = +0.050$ (Appendix Table 18). Structure-family modules partition encoders on a different axis from R1 and R2: H1 spatial relations is object-local for CLIP and DINOv2-reg, borderline object-local for DeiT and DINOv3 ($\Delta_{\text{loc}} \in [+0.056, +0.062]$), and image-wide for SigLIP 2 and MAE (Appendix Table 19); the extended panel adds BioCLIP as a contrastive counter-example, so pretraining objective alone does not explain H1 locality (Appendix A.6). H2 counting is most strongly object-local on CLIP and SigLIP 2 (Appendix Table 20); DINOv2-reg is the only encoder below threshold under the dual-nuisance baseline (small-crop OOD; §7). H3 distractor robustness is object-local on five of six base encoders by target-vs-distractor locality, with MAE the image-wide exception. H3 partitions differently from H1: SigLIP 2 is image-wide on H1 but object-local on H3, indicating that Structure is not a single scalar property but a set of claim-specific tests.

**R1 position: three regimes and three reversals.** Position metrics under the three pools fall into three regimes (Table 1): object-local (CLIP-B/16, DINOv2-reg), image-wide (MAE, DeiT), and patch-position dominated (DINOv3, SigLIP 2). The two encoders in each regime differ in pretraining objective, positional encoding, and supervision modality, so the panel does not disentangle which factor drives the regime. We compute seven readouts grouped into four diagnostic axes; per-metric labels (Table 4) are the protocol's output, and definitions are in Appendix A.2.

The 49-way readout predicts which of the 49 object-placement cells contains the object. Figure 4 shows the same reversal across the six-base panel: the blank same-footprint pool — the same probe trained on a uniform-grey image, pooled at the patch indices the object would have occupied — reaches near-perfect 49-way object-position decoding for every encoder, often exceeding the object pool. Blank balanced accuracy is at least $0.978$ across all eight audited representations and survives shuffled-label and footprint-count metadata-oracle controls over five seeds (Appendix A.12). The SigLIP 2 reversal is one instance of a broad coordinate-readout effect, not a one-model anomaly.

For SigLIP 2, the matched-background pool nonetheless drops to $0.581$, so a real object-vs-non-object distinction survives that the blank-canvas reversal alone would obscure ($\Delta_{\text{loc}} = +0.410$ on 49-way bal-acc). Two further reversals appear on the Jaccard remapping metric in Table 1. DINOv3's blank-canvas Jaccard ($0.532$) exceeds its object-pool Jaccard ($0.400$): a uniform grey forward produces patch-token features whose active set varies strongly with grid position, and under this metric the scaffold-only signal exceeds any object-incremental signal on DINOv3. MAE's object-pool Jaccard ($0.431$) versus blank ($0.017$) gives the largest $\Delta_{\text{inc}}$ we observe ($+0.414$), but matched-bg also reaches $0.432$, so $\Delta_{\text{loc}} \approx 0$: MAE produces strong object-induced patch geometry that propagates image-wide rather than object-pixel-localised.

**Two modules fail their design criteria, and the controls catch both.** An initial R2 design (400 LVIS objects, three scales, fixed centre position) was rejected by the patch-count baseline: object-pool scores landed within three percentage points of a patch-count predictor at 0.975 (Appendix Table 8). A footprint-fixed redesign holds patch count constant across scales; the patch-count baseline then sits at chance, every encoder clears the $\Delta_{\text{inc}} \geq 0.20$ threshold, and MAE is the only

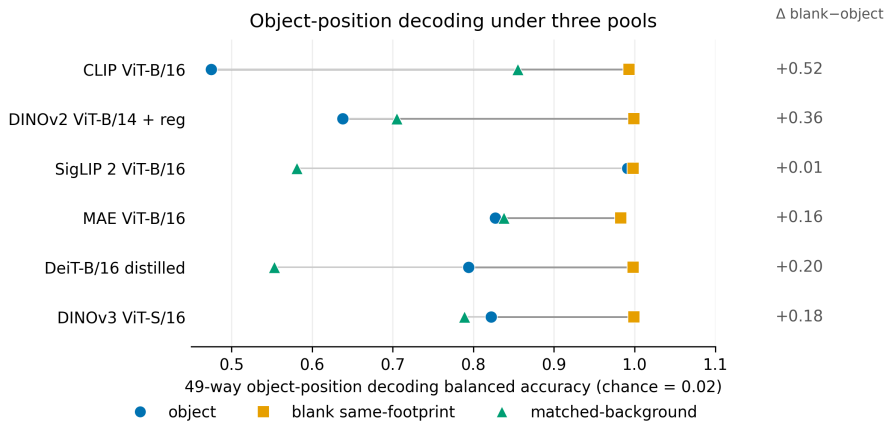


Figure 4: **Object-position decoding under three pools.** A probe predicts which of 49 object-placement cells contains the rendered LVIS object. Across all six base encoders, the blank same-footprint pool, which uses a uniform-grey image but pools the same patch indices as the object footprint, reaches near-ceiling balanced accuracy. High object-pooled position decoding can therefore be explained by patch-coordinate information exposed by the pooling rule rather than by object-local visual evidence.

encoder labelled object-local on the redesigned R2 (Appendix Table 18). An initial H2 counting design saturated the footprint-area baseline at 1.000, sending every encoder below the dual-nuisance threshold (Appendix Table 9). A target-area-fixed redesign holds total alpha-mask area constant across counts via per-$N$ object sizes $s_N = \lfloor \sqrt{T/N} \rfloor$; the footprint-area baseline drops to 0.382, five of six encoders clear the $+0.15$ dual-nuisance threshold, and CLIP and SigLIP 2 are the most strongly object-local on counting (Appendix Table 20). The redesign criterion is removal of the saturating nuisance baseline, not optimisation against the encoder panel.

**SAE bases.** Two SAE bases (DINOv3-SAE and BioCLIP-SAE) have full three-pool coverage on the v2 modules under persisted activation-mean statistics; both show object-pool above matched-background by margins consistent with the partition the raw bases produce, with the non-localised-features warning of Han et al. [16] accounted for by reporting reconstruction calibration alongside causal scores (Appendix A.1). Two older SAE bases (CLIP-SAE, DINOv2-reg-SAE) have R1 three-pool numbers but incomplete v2 blank-pool coverage under the older preprocessing path; their v2 results are reported as method observations in Appendix A.5 rather than as headline regime assignments.

**Transfer to natural images.** We apply an analogous three-pool readout to natural-image LVIS object boxes (object box, same patch indices on a uniform-grey forward, and matched-background non-object patches in the same image) and score balanced-accuracy category decoding per pool. Five of six base encoders show object pool above the strongest control by $+0.064$ to $+0.204$; MAE inverts to $-0.047$ under matched-background domination (Appendix A.8, Figure 5). The per-encoder partition matches the synthetic-R1 partition — MAE is the same matched-background-dominated outlier — so the regime structure is not an artefact of the grey-canvas protocol. The bridge readout is smaller in scale and uses category decoding rather than $\Delta_{\text{inc}}/\Delta_{\text{loc}}$, so it is a transfer check, not a replacement for the controlled three-pool modules: it tests whether the encoder partition transfers, not whether each synthetic module's regime transfers.

## 5 Readout and intervention methods

The method axis evaluates probing and intervention methods on the same modules and audited representations. Each method is scored on four axes following Mueller et al. [29]: faithfulness against controlled-probe evidence, specificity against cross-module interference, generalisation across encoders, and concreteness. We evaluate M1 (naive linear probe), M2 (controlled linear probe), and

| Representation | object | blank | matched-bg | $\Delta_{\text{inc}}$ | $\Delta_{\text{loc}}$ | regime |
|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.288 | 0.183 | 0.163 | $+0.105$ | $+0.125$ | object-local |
| DINOv2-reg raw | 0.408 | 0.155 | 0.204 | $+0.253$ | $+0.205$ | object-local |
| DeiT-B/16 raw | 0.341 | 0.251 | 0.311 | $+0.090$ | $+0.030$ | image-wide |
| MAE-B/16 raw | 0.431 | 0.017 | 0.432 | $+0.414$ | $-0.001$ | image-wide |
| DINOv3-S/16 raw | 0.400 | 0.532 | 0.541 | $-0.133$ | $-0.142$ | patch-position |
| SigLIP 2-B/16 raw | 0.123 | 0.164 | 0.150 | $-0.041$ | $-0.027$ | patch-position |

Table 1: **R1 position regimes on Jaccard remapping $R^2$ under three pools, six base encoders.** All cells are Jaccard remapping $R^2$: the linear-fit coefficient of determination relating pairwise representational distance to grid-cell distance over the active-set Jaccard overlap of pooled feature vectors (full definition in Appendix A.2). Three regimes appear: object-local (CLIP, DINOv2-reg), image-wide (MAE, DeiT), and patch-position dominated (DINOv3, SigLIP 2). The 49-way balanced-accuracy reversal is in Figure 4; SAE-basis rows in Appendix A.5; per-metric R1 detail and borderline entries in Appendix A.2. Cells within a regime differ in pretraining objective, positional-encoding scheme, and supervision modality.

M3 (top-$k$ subspace probe) across the six-encoder raw panel, and six cached M4 (SAE feature ablation under plausible replacement) cases across CLIP-SAE and DINOv3-SAE.

M3 (top-$k$ subspace probe) reports a separate concreteness score $k_{50}$; we move its values and discussion to Appendix A.3 because $k_{50}$ does not enter the regime taxonomy.

**M4 SAE feature ablation under plausible replacement.** M4 ranks SAE features by per-feature module-axis SS, ablates the top-$k = 50$ under three plausible-replacement controls (mean, noise, shuffle), and reports the maximum drop as the causal score; a random-feature control matched in $k$ measures specificity. Zero ablation enters only as an out-of-distribution comparator [34]: on DINOv3-SAE H1, zero ablation drops the same fifty features by $0.322$ versus a plausible-replacement drop of $0.162$, illustrating the magnitude of the OOD gap that motivates the plausible-replacement comparator. Across six cached cases (CLIP-SAE: W3, H1, R2; DINOv3-SAE: W3, H1, H2), five faithfulness margins (plausible drop minus random-feature drop) are positive ($+0.664$ CLIP-SAE W3; $+0.309$, $+0.161$ DINOv3-SAE W3, H1; $+0.087$ CLIP-SAE H1; $+0.025$ CLIP-SAE R2) and one is negative ($-0.005$ DINOv3-SAE H2). M4 is interpreted only when the SAE basis itself reconstructs the audited representation faithfully: the DINOv3-SAE rows have reconstruction cosine above $0.98$ on the same caches (Appendix Table 3), so the DINOv3-SAE W3 and H1 rows carry the main causal interpretation; CLIP-SAE rows are reported as method observations.

## 6 Task-conditioned downstream alignment

Downstream validation is a stress test of whether the claim-conditioned evidence profiles in §4–§5 help choose encoders for particular task families. Five frozen-encoder tasks span the three claim families: ADE20K-150 segmentation (Where), NYUv2 depth (geometry-heavy Where), iNat fine-grained classification (What), RefCOCO+ grounding (Structure), and FSC-147 object counting (Structure). Each task uses the same shallow linear-head architecture across encoders. We report Spearman rank correlation between the suite's selected ranking and the downstream ranking, with the naive object-pool linear probe (M1) as the comparator. A locked machine-readable registry pairs each task with its claim-conditioned selector: R1 object-locality for spatial readout, R2+H1 object-locality for geometry-heavy depth, W1+W3 object-locality for fine-grained classification, and H2 counting locality for FSC-147.

Table 2 reports per-task gaps. D3 RefCOCO+ is reported as candidate-region evidence supporting D1's positive segmentation result rather than as a fifth headline gap; four of six encoders beat the strongest oracle (area prior $0.496$, SigLIP 2 at $0.547$), but the selected gap sign-flips when MAE is removed (Appendix A.11).

At $n$=6 encoders the rank correlations are sensitive to which encoders enter the panel; we therefore treat the gaps as diagnostic alignment checks rather than confirmatory rankings. D1 ADE20K is the only outcome that is both positive and stable under expansion to $n$=8. D2 sign-flips under expansion, D4 is negative (identity and colour content is image-wide on every encoder, so the W1+W3

| Task | Family | Selector | Suite $\rho$ | Naive $\rho$ | Gap |
|---|---|---|---|---|---|
| D1 ADE20K seg | Where | `where_r1_loc.` | $+0.086$ | $-0.200$ | $\mathbf{+0.286}$ |
| D2 NYUv2 depth | Where+Structure | `geometry_r2_h1_loc.` | $+0.200$ | $-0.618$ | $\mathbf{+0.818}$ |
| D4 iNat-mini cls | What | `what_w1_w3_loc.` | $-0.257$ | $+0.314$ | $\mathbf{-0.571}$ |
| D5 FSC-147 count | Structure | `structure_h2_loc.` | $-0.657$ | $-0.657$ | $\mathbf{0.000}$ |

Table 2: **Downstream alignment rank correlation across four cached-feature downstream tasks at $n = 6$ base encoders.** Locked task-family selectors, five-seed median downstream summaries. D2 is positive across seeds and leave-one-out at $n = 6$ but sign-flips under panel expansion to $n = 8$ (Appendix A.10). D1 is positive across seeds and sensitive to leave-one-out. D4 is negative across seeds. D5 is near zero with mixed seed signs. D3 RefCOCO+ uses a candidate-region setup with selected gap $+0.286$ on the full panel and is reported in the appendix because the gap sign-flips when MAE is removed (Appendix A.11). Per-task seed, leave-one-out, and panel-expansion ranges are in Appendix A.10.

*locality* selector is misaligned with category-prototype-driven ranking), and D5 is near zero on both panels. Per-task seed, leave-one-out, and panel-expansion ranges are in Appendix A.10.

# 7 Scope and reproducibility

**Intended use.** Sparse Groundings is for falsifying object-local representation claims in spatial token maps, comparing readout regimes across encoders and SAE bases under controlled synthetic stimuli, and testing whether an interpretability intervention's effect survives nuisance-matched controls. It is not a ranking of general-purpose vision encoders, a predictor of VLM downstream accuracy, a causal claim about how features are used by downstream systems, or a safety or deployment certification.

**Scope.** Stimuli are objects on a uniform-grey $224\times224$ canvas with $7\times7$ object-centre placements: a stress test for claim falsification under known nuisance variables, not an approximation to natural-image deployment. The natural-image bridge (§4, Appendix A.8) is a transfer check. The encoder panel is six raw bases plus two SAE bases; sensitivity to leave-one-encoder-out and panel expansion to $n{=}8$ is in Appendix A.10. The protocol does not separate which factor (pretraining objective, positional encoding, supervision modality) drives a given regime; coverage gaps (continuous-jitter centres, alternative pooling, non-grey blank controls, held-out LVIS subsets, footprint jitter, cutout) are listed in Appendix A.12.

**Compute footprint.** Producing the submitted empirical receipts required approximately 25–40 GPU-hours on a workstation with two NVIDIA RTX 4090 GPUs (24 GB VRAM each), plus $\sim$ 20–40 CPU-hours for probe fitting, aggregation, and figure/table generation. Cached-feature artefacts occupy $\sim$ 3 GB; regenerating tables, figures, and JSON receipts from caches takes a few CPU-minutes with no GPU forwards.

Sparse Groundings turns claims of the form "the encoder represents $X$" into per-claim records that survive or fail named controls. The object-pooled linear probe, currently the default audit primitive for vision-encoder evaluation, can be confounded by patch-coordinate readout from the pooling rule itself. Different claim families behave differently under the same controls, and downstream alignment is task-conditioned. The released artifact (`pip`-installable library, CLI, evaluation cards, locked selector registry, schema-validated claim cards, and a deterministic reproduction manifest; Appendix A.9) replaces a global score with a claim-level audit on a chosen encoder and module.

# References


[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[2] Emmanuel Ameisen et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, 2025.

[3] Samyadeep Basu, Matthew Grayson, Cheyenne Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2024.

[4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[5] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Thomas Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

[6] Wenxiao Cai, Yaroslav Ponomarenko, Jiaming Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

[7] 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*, pages 9650–9660, 2021.

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

[10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, 2024.

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

[12] Mengfei Du, Bo Wu, Zhenfang Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 2024.

[13] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, William Yang Wang Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024.

[14] Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11462–11482, 2025. doi: 10.18653/v1/2025.naacl-long.571.

[15] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019.

[16] Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision. *arXiv preprint arXiv:2509.00749*, 2025.

[17] 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*, pages 16000–16009, 2022.

[18] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

[19] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

[20] Stephen Joseph et al. Prisma: An open source toolkit for mechanistic interpretability in vision and video. In *CVPR Workshop on Mechanistic Interpretability for Vision*, 2025.

[21] Stephen Joseph et al. Steering CLIP's vision transformer with sparse autoencoders. In *CVPR Workshop on Mechanistic Interpretability for Vision*, 2025.

[22] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

[23] Seongyun Kang, Junyoung Kim, Joon-Young Kim, and Seong Joon Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9339–9350, 2025.

[24] Seongyun Kang, Junyoung Kim, Joon-Young Kim, and Seong Joon Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, 2025.

[25] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677, 2018.

[26] Byeong Jo Kim, Hyeyeon Choi, Hyeonah Jang, and Sang Woo Kim. Understanding gaussian attention bias of vision transformers using effective receptive fields. In *British Machine Vision Conference*, 2023.

[27] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, 2025. arXiv:2412.05276.

[28] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

[29] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45069–45108, 2025.

[30] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023.

[31] 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.

[32] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

[33] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations*, 2023.

[34] Felipe Parodi et al. Zero-ablation overstates register function in DINO vision transformers, 2026.

[35] Jiaxing Qi, Jun Liu, Hao Tang, and Zhihong Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models, 2026. OpenReview preprint.

[36] 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[37] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128, 2021.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[39] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023.

[40] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo José, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[41] Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G. Campolongo, Chan Hee Song, David E. Carlyn, Li Dong, Wasila M. Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024.

[42] Samuel Stevens, Jaden Beattie, Tanya Berger-Wolf, and Yu Su. Towards open-ended visual scientific discovery with sparse autoencoders. *arXiv preprint arXiv:2511.17735*, 2025.

[43] Samuel Stevens, Wei-Lun Chao, Tanya Berger-Wolf, and Yu Su. Sparse autoencoders for scientifically rigorous interpretation of vision models. *arXiv preprint arXiv:2502.06755*, 2025.

[44] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermyn, Tom Henighan, Shan Carter, Robert Cunningham, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024.

[45] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10599–10609, 2025.

[46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

[47] 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*, pages 9568–9578, 2024.

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357, 2021.

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

[50] Jingyun Wang and Guoliang Kang. Learn to rectify the bias of CLIP for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4102–4112, 2024.

[51] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, Guang Shi, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[52] 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*, 2023.

[53] Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. *arXiv preprint arXiv:2502.20578*, 2025.

[54] Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 11591–11609, 2025.

# A Appendix

## A.1 Reconstruction calibration for SAE bases

The SAE calibration receipt checks whether the SAE basis reconstructs the paired raw patch features on the same synthetic caches used by the module and M4 receipts. DINOv3-S/16-SAE has complete pooled-cache calibration on W3, R2, H1, and H2, with reconstruction cosine above 0.98 and explained variance near 0.88–0.90. BioCLIP-SAE also has complete rows on these v2 modules, but its explained variance is only 0.44–0.49 despite reconstruction cosine above 0.91, so we treat it as a controlled representation row rather than main M4 causal support. The older CLIP-SAE and DINOv2-reg-SAE caches provide exact explained variance where paired raw / SAE caches exist, but exact reconstruction cosine and MSE are blocked because the old extractor did not persist the per-run activation mean.

The decoded-delta base-space cross-check reuses the existing M4 receipts without changing their SAE-space values. The top three positive receipts remain positive after decoded SAE feature deltas are substituted into base-encoder caches: CLIP-SAE W3 has base-space faithfulness margin $+1.497$ (SAE-space $+0.664$), DINOv3-SAE W3 has $+0.172$ (SAE-space $+0.309$), and DINOv3-SAE H1 has $+0.107$ (SAE-space $+0.161$). This is a consistency check, not an equivalence proof or a full model forward pass.

## A.2 Per-metric regimes for the eight panel representations

R1 reports seven readouts in four axes: *organisation* (Jaccard remapping $R^2$ on full and top-decile feature sets, cosine remapping $R^2$); *availability* (ridge $R^2$ for $g_x$, $g_y$ at $\alpha = 1$ and 49-way balanced accuracy on a held-out 80/20 split, seed 42); *per-feature dominance* (*frac_position*$_i$ =

| SAE basis | stimulus | recon. cos. | EV | mean $L_0$ |
|---|---|---|---|---|
| DINOv3-S/16 SAE | W3 colour | 0.986 | 0.892 | 478 |
| DINOv3-S/16 SAE | R2 scale-v2 | 0.987 | 0.894 | 490 |
| DINOv3-S/16 SAE | H1 relations | 0.989 | 0.896 | 574 |
| DINOv3-S/16 SAE | H2 counting-v2 | 0.988 | 0.879 | 514 |
| BioCLIP-B/16 SAE | W3 colour | 0.916 | 0.459 | 1943 |
| BioCLIP-B/16 SAE | R2 scale-v2 | 0.919 | 0.443 | 1993 |
| BioCLIP-B/16 SAE | H1 relations | 0.934 | 0.491 | 4301 |
| BioCLIP-B/16 SAE | H2 counting-v2 | 0.923 | 0.463 | 2912 |

Table 3: SAE reconstruction calibration on pooled synthetic caches. EV is explained variance. CLIP-SAE and DINOv2-reg-SAE rows from the older extractor are omitted from the table because reconstruction cosine and MSE are blocked by missing activation-mean statistics.

$\mathrm{SS}_{\text{position},i}/\mathrm{SS}_{\text{total},i}$ via a balanced two-way object $\times$ position decomposition); and *concentration* ($k_{50}$, the smallest $k$ whose top features by $\mathrm{SS}_{\text{position}}$ explain 50% of total). The three regimes are not uniform across metrics: CLIP and DINOv2-reg are object-local on Jaccard and cosine but patch-position dominated on the linear-decoding metrics, since the patch-coordinate scaffold saturates linear position decoding for every representation.

| Representation | Jac. full | Jac. top-10% | Cosine | Ridge $R^2_x$ | Ridge $R^2_y$ | Bal-acc 49 |
|---|---|---|---|---|---|---|
| *Base representations* | | | | | | |
| CLIP-B/16 raw | **OL** | **OL** | **OL** | PP | PP | PP |
| DINOv2-reg raw | **OL** | **OL** | **OL** | PP | PP | PP |
| SigLIP 2 raw | PP | **OL** | PP | PP | PP | PP |
| DINOv3-S/16 raw | PP | **OL** | PP | PP | PP | PP |
| MAE-B/16 raw | IW | IW | IW | PP | PP | PP |
| DeiT-B/16 raw | IW | PP | **OL** | PP | PP | PP |
| *SAE bases (over the matching base representation)* | | | | | | |
| CLIP-B/16 SAE | inc | inc | inc | inc | inc | inc |
| DINOv2-reg SAE | inc | inc | inc | inc | inc | inc |

Table 4: Per-metric regime label per representation. Codes: **OL** = object-local, IW = image-wide, PP = patch-position dominated readout, inc = incomplete (SAE blank pool not yet under the per-run-statistics fix). Tolerance $\varepsilon = 0.05$. All six raw representations land in PP on the linear-decoding metrics (Ridge $R^2_{x/y}$, Bal-acc 49) because the patch-coordinate scaffold saturates linear position decoding for every representation. The remapping metrics (Jaccard and cosine) carry the cross-representation variation.

### A.3 Top-$k$ candidate intervention targets

$k_{50}$ is the smallest $k$ for which the top-$k$ features by per-feature SS along the relevant axis recover 50% of the full-feature probe's balanced accuracy. The same construction is applied to position-axis SS (R1, Table 6) and module-axis SS across the four v2 modules (Table 5). Smaller $k_{50}$ means a more concentrated subspace and a smaller candidate intervention target set. $d_{\text{features}} = 768$ for the ViT-B class and $384$ for DINOv3-S/16.

### A.4 V1 module result tables

The W1 instance, R2 v1 scale, and H2 v1 counting result tables are placed here in the appendix because the v1 R2 and v1 H2 designs both saturate a nuisance baseline; the corrected v2 designs are the main-text result tables. W1 is at near-uniform 1.000 on the object pool and is included as the controls-survive-broadcast example.

**W1 instance recognition.** Multinomial logistic on per-stimulus pooled features predicting one of 400 `object_id` labels under position-held-out 80/20 split. Chance $= 1/400 = 0.0025$.

| Representation | W3 colour | R2 scale-v2 | H1 relations | H2 counting |
|---|---|---|---|---|
| CLIP-B/16 raw | 2 | 2 | 18 | 1 |
| DINOv2-reg raw | 5 | 1 | 18 | 1 |
| SigLIP 2-B/16 raw | 2 | 1 | 3 | 1 |
| DINOv3-S/16 raw | 8 | 1 | 12 | 1 |
| MAE-B/16 raw | 2 | 2 | 8 | 1 |
| DeiT-B/16 raw | 3 | 1 | 8 | 1 |

Table 5: M3 top-$k$ subspace $k_{50}$ per (representation, module). W3 colour and R2 scale-v2 concentrate to $k_{50} \leq 8$; H1 relations distributes ($k_{50} \in [3, 18]$); H2 counting sits at near-chance object-pool accuracy (Table 9) so $k_{50} = 1$ is uninformative.

| Representation | object pool | blank pool | matched-bg pool |
|---|---|---|---|
| *Base representations* | | | |
| CLIP-B/16 raw | 6 | 1 | 2 |
| DINOv2-reg raw | 38 | 164 | 138 |
| SigLIP 2 raw | 2 | 2 | 56 |
| DINOv3-S/16 raw | 30 | 32 | 47 |
| MAE-B/16 raw | 122 | 17 | 121 |
| DeiT-B/16 raw | 7 | 30 | 119 |
| *SAE bases (over the matching base representation)* | | | |
| CLIP-B/16 SAE | 5 | – | 3 |
| DINOv2-reg SAE | 31 | – | 64 |

Table 6: R1 $k_{50}$ per (representation, pool). CLIP and SigLIP 2 concentrate position into very few features on the object pool ($k_{50} = 6$ and 2); MAE spreads to $k_{50} = 122$. Encoders where the matched-bg subspace is wider than the object subspace (CLIP, SigLIP 2, DeiT) carry distributed non-object spatial structure.

**R2 v1 scale (centre-position only, three scales).** Multinomial logistic on the 3-class scale label, object-held-out 80/20 split, chance $= 0.333$. Patch-count baseline predicts scale from $|\mathcal{F}_{\text{obj}}(x)|$ alone via logistic regression on the scalar footprint count.

**H2 v1 counting (single-area object bank).** $N$ objects ($N \in \{1, \ldots, 6\}$) on canvas, 50 random configurations per $N$ (300 stimuli per pool). Object-held-out scene split, $n_{\text{test}} = 60$, chance $= 1/6 = 0.167$. Footprint-area baseline predicts count from total alpha-mask pixel area.

### A.5 SAE locality on the v2 module panel

CLIP-SAE moves $\Delta_{\text{loc}}$ *into* the object-local regime ($\geq \varepsilon = 0.05$) on three of four v2 modules where raw CLIP was image-wide. DINOv2-reg-SAE moves $\Delta_{\text{loc}}$ by at most $+0.04$ on H1 and the wrong direction on H2 v2 ($-0.073$, worsening the image-wide-with-flip already present in raw).

### A.6 Extended representation panel (v2 modules)

The extended panel adds DINOv2 no-reg and BioCLIP raw plus saev DINOv3-S/16 24K and saev BioCLIP-B/16 24K to the main six-encoder plus two-SAE-basis panel, for twelve audited representations. R1 position is not reported on the four new rows because the LVIS-400 position-grid features are not cached for them; W3, R2, H1, and H2 v2 are. The headline $n$=6 claims and the framing in §3 remain anchored to the six-base panel. Two of the four SAE bases (saev CLIP, saev DINOv2-reg) carry an `incomplete` regime label because the older saev preprocessing path's center/scale statistics push the SAE encode off-distribution on a degenerate blank-only forward; for those rows we report $\Delta_{\text{loc}}$ from object and matched-background only.

**H2 v2 small-crop OOD on DINOv2.** The H2 v2 redesign renders objects at side lengths 20–48 px; both DINOv2-reg and DINOv2 without registers fail the dual-nuisance DoD (family-level not register-specific) while every other encoder clears it by $+0.4$ or more, consistent with DINOv2's

| Representation | object | blank | matched-bg | $\Delta_{\text{inc}}$ | $\Delta_{\text{loc}}$ |
|---|---|---|---|---|---|
| CLIP-B/16 raw | 1.000 | 0.002 | 0.995 | +0.998 | +0.005 |
| DINOv2-reg raw | 1.000 | 0.001 | 0.998 | +0.999 | +0.002 |
| DINOv3-S/16 raw | 1.000 | 0.004 | 1.000 | +0.996 | 0.000 |
| SigLIP 2-B/16 raw | 1.000 | 0.003 | 0.995 | +0.997 | +0.005 |
| MAE-B/16 raw | 1.000 | 0.001 | 1.000 | +0.999 | 0.000 |
| DeiT-B/16 raw | 1.000 | 0.002 | 1.000 | +0.999 | 0.000 |

Table 7: W1 instance recognition under position-held-out generalisation. Object and matched-background pools recover near-perfect per-instance recognition; blank canvas is at chance.

| Representation | object | blank | matched-bg | patch-count | object - patch-count |
|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.979 | 0.996 | 0.988 | 0.975 | +0.004 |
| DeiT-B/16 raw | 0.992 | 1.000 | 0.983 | 0.975 | +0.017 |
| DINOv2-reg raw | 0.967 | 0.996 | 0.988 | 0.979 | −0.012 |
| DINOv3-S/16 raw | 0.996 | 1.000 | 0.996 | 0.975 | +0.021 |
| MAE-B/16 raw | 0.988 | 0.979 | 0.983 | 0.975 | +0.013 |
| SigLIP 2-B/16 raw | 1.000 | 1.000 | 0.946 | 0.975 | +0.025 |

Table 8: R2 scale-v1. Object-pool scores within $\pm 3$ pp of the patch-count baseline; blank and matched-background controls are also near ceiling. At three coarse scales rendered at a fixed centre position, the scale label is almost recoverable from the footprint count alone. This v1 negative motivated the v2 footprint-fixed redesign (Appendix Table 18).

dense-pretraining objective targeting full-bbox objects. An H2 v2.5 design varying object size jointly with $N$ across a wider range would settle this.

**W3 cross-representation spread DoD unreachable on grey backgrounds.** An earlier W3 success criterion targeted at least two encoders with matched-bg pool below half the object pool. On grey-background single-object stimuli every encoder's matched-bg exceeds 0.85, so the spread is unreachable. Cluttered backgrounds (H3-style) or multi-object scenes with one target colour are the candidate redesigns.

### A.7 Encoder panel coverage

The measured panel is twelve audited representations: eight base encoders and four SAE bases. Six base rows and two SAE rows have R1 + v2 module coverage; two base rows and two SAE rows have v2 only. All ViT-class representations use block 10 patch tokens with CLS and registers stripped before pooling. SAE hook points match the checkpoints' training extraction (`data.layer:-2 = block 10` for both 12-block backbones).

### A.8 Natural-image bridge

We apply the three-pool readout to a natural-image LVIS subset: each example contributes an object-box pool, a blank-coordinate pool that uses the same patch indices on a uniform-grey forward, and a matched-background pool drawn from non-object patches in the same image. We score balanced-accuracy category decoding from each pool with a frozen-feature linear head and report the object-pool minus the strongest control as a per-encoder margin.

### A.9 Released artifact components

The minimal command-line workflow is `sg run --cell <name> --module <m> --method <M#>`, which returns a claim card listing the object, blank, and matched-background scores, the two deltas, the regime label, the module's nuisance baseline, and the JSON receipt path that produced the entry. The released artifact ships a `pip`-installable library and CLI; per-module evaluation cards under `paper/eval-cards/`; 73 schema-validated claim cards covering the audited entries; the locked selector registry that pairs each downstream task with its claim-conditioned selector; and a cached-feature reproduction manifest hashing 105 source and gen-

| Representation | object | blank | matched-bg | fp-area | $\Delta_{\text{loc}}$ | $\Delta_{\text{inc-dual}}$ |
|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.515 | 0.212 | 0.558 | 1.000 | $-0.043$ | $-0.485$ |
| DINOv2-reg raw | 0.447 | 0.176 | 0.609 | 1.000 | $-0.162$ | $-0.553$ |
| SigLIP 2-B/16 raw | 0.531 | 0.213 | 0.439 | 1.000 | $+0.092$ | $-0.469$ |
| DINOv3-S/16 raw | 0.449 | 0.183 | 0.558 | 1.000 | $-0.109$ | $-0.551$ |
| MAE-B/16 raw | 0.626 | 0.171 | 0.483 | 1.000 | $+0.144$ | $-0.374$ |
| DeiT-B/16 raw | 0.529 | 0.180 | 0.468 | 1.000 | $+0.061$ | $-0.471$ |

Table 9: H2 counting v1. Footprint-area baseline saturates at 1.000 across the panel; the LVIS-20 object bank used here has small per-object alpha-mask area variance, so total area $\approx N \cdot \overline{a}$. Every representation's object pool sits well below the dual-nuisance ceiling, so $\Delta_{\text{inc-dual}}$ is large and negative for every representation. This v1 design failure motivated the v2 target-area-fixed redesign (Appendix Table 20).

| $\Delta_{\text{loc}}$ | W3 colour | R2 scale-v2 | H1 relations | H2 counting v2 |
|---|---|---|---|---|
| *CLIP base → saev CLIP-B/16 24K basis* | | | | |
| CLIP-B/16 raw (base) | $+0.018$ | $-0.031$ | $+0.050$ | $+0.256$ |
| CLIP-B/16 SAE | $+0.378$ | $+0.131$ | $+0.219$ | $+0.255$ |
| *DINOv2-reg base → saev DINOv2-reg ViT-B/14 24K basis* | | | | |
| DINOv2-reg raw (base) | $-0.054$ | $-0.037$ | $+0.062$ | $-0.080$ |
| DINOv2-reg SAE | $-0.037$ | $-0.037$ | $+0.100$ | $-0.153$ |

Table 10: SAE-vs-base $\Delta_{\text{loc}}$ on the v2 modules. Each SAE row is a property of the basis trained over the matching base representation, not a property of the underlying encoder. CLIP-SAE pulls content into object-pool patches on attribute, relations, and counting; on the one module where raw CLIP was already object-local (H2 v2), the SAE preserves the locality. DINOv2-reg-SAE preserves its raw R1 image-wide-with-flip behaviour on the v2 panel.

erated artefacts. Matched-background indices are byte-reproducible from a `blake2b` digest of $(\textit{object_id}, g_x, g_y, \textit{scale}, \textit{patch_size}, \textit{base_seed})$.

### A.10 Validation seed stability and sign sensitivity

At $n = 6$ non-parametric bootstrap CIs are wide; we report five-seed median gaps and sign stability instead. Two sensitivity dimensions: seed gap range across five head seeds, and leave-one-out gap range across the six base encoders. D1 and D2 are seed-stable positive, D4 stable negative, D5 sign-fragile. Under leave-one-out, only D2 stays positive across all six drops; D1 sign-flips when MAE is dropped. A joint sign check over D1/D2/D4 gives $3/3$ matching directions ($p = 0.25$ under a two-sided binomial check); this is descriptive, not a power claim.

**Expanded-panel sensitivity ($n = 8$).** Adding BioCLIP raw and DINOv2 no-reg raw under the locked selector registry, $n_{\text{cells actual}} = 8$ for every selector with no rows filtered. D1 ADE20K strengthens ($+0.286 \rightarrow +0.381$). D2 NYUv2 flips sign ($+0.818 \rightarrow -0.210$, sign-stable across five seeds). D4 iNat attenuates ($-0.571 \rightarrow -0.132$). D5 stays sign-fragile. Selectors, task definitions, and $\varepsilon$ are locked across panels, so this is panel conditioning rather than selector drift.

### A.11 Validation hardening receipts

**D3 RefCOCO+ candidate-region receipt.** D3 scores 768 expressions $\times$ 6 candidates each (ground-truth box, two jittered positives, two same-area distractors, one random) against centre, area, and random oracles. The strongest oracle is the area prior at Acc@0.5 $= 0.496$; four of six encoders beat it (Table 15), SigLIP 2 highest at $0.547$. Under the locked `object_local_w1_w3_h3_target` selector, the suite-vs-naive $\rho$-gap is $+0.286$ on the full six-base panel. Five-seed scorer agreement is by construction (convex lbfgs over fixed candidates), not a stability claim. Leave-one-out flips sign when MAE is removed ($+0.286 \rightarrow -0.100$); D3 therefore remains appendix-only pending candidate bootstrap or panel expansion.

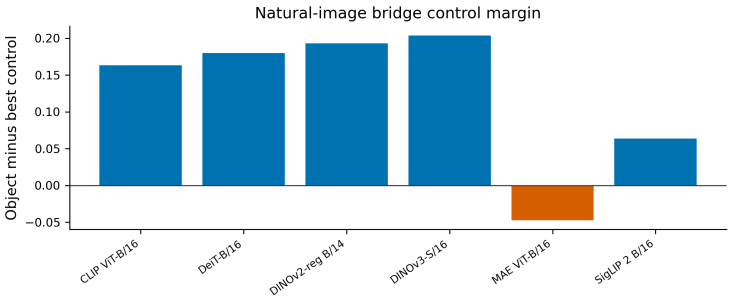


Figure 5: **Natural-image bridge: object-pool margin over the strongest control.** Balanced-accuracy category-decoding margin on natural-image LVIS object boxes per base encoder. Five of six encoders show positive margins (DINOv3 $+0.204$, DINOv2-reg $+0.193$, DeiT $+0.180$, CLIP $+0.163$, SigLIP 2 $+0.064$); MAE inverts to $-0.047$ under matched-background domination, the same outlier as on synthetic R1. The bridge is intended as a transfer check, not a replacement for the controlled three-pool modules.

**External-benchmark scope.** MMVP is reported as an appendix-only internal image-pair separability comparison over 150 pairs (not VQA accuracy); the selector-derived suite ranking has $\rho = -0.143$ against mean separability over the six-base panel. VSR, What's Up, and BLINK are not reported because local per-encoder score files are unavailable.

### A.12 Control-hardening receipts

Deterministic blank-pool informativeness passes: high blank scores indicate coordinate / scaffold availability rather than object content. Random-Gaussian features do not create object-local support; a position-only null recovers scaffold labels and anchors the patch-position regime. R1 49-way blank balanced accuracy reaches $\geq 0.978$ across all eight audited representations (including CLIP-SAE 0.978 and DINOv2-reg-SAE 0.998 under persisted activation-mean stats). W3 colour, H3 distractor robustness, and R1 49-way position all pass five-seed shuffled-label and footprint-count metadata-oracle selectivity across the six-base panel; R1 covers the object, blank same-footprint, and matched-background pools, with overall verdict `pass_probe_selectivity` (Table 16). H4 binding has six-base object, blank, matched-background, and distractor receipts with marginal oracles at chance, but every encoder is image-wide under the matched-background control, so background-token features recover the binding label. The natural-image sanity check has six-base pooled-box caches; object boxes beat controls for most encoders, while MAE is matched-background dominated. Blank intensity / noise / texture variants, held-out LVIS subsets, footprint jitter, alternative pooling, and cutout remain unevaluated.

### A.13 Per-module v2 result tables

### A.14 Extended related work

**Visual perception failures and vision-centric MLLMs.** Tong et al. [47] construct CLIP-blind pairs and the MMVP benchmark; Cambrian-1 [46] shows vision-encoder choice dominates MLLM performance on CV-Bench; BLINK [13] reformats 14 vision tasks as multiple-choice (humans 95.7%, GPT-4V 51.3%); MMStar [9] shows many VLM benchmarks do not require vision content; Yuksekgonul et al. [52] introduce ARO and show contrastive VLMs behave like bags-of-words on relation and word-order tasks.

**Spatial benchmarks for VLMs and VLM internals.** Behavioural spatial benchmarks evaluate end-to-end VLM accuracy: [22, 28, 51, 54]. Spatial-VLM training and grounding work: [6, 8, 10, 12]. VLM-internals work attributes specific failures to position bias [45], attention sinks [24], norm imbalance [35], and localisation-relevant attention heads [14, 23].

**Visual explanations and the falsifiability discipline.** Classic visual interpretability methods include [4, 7, 25, 30, 38]. Adebayo et al. [1] show several saliency methods produce similar expla-

nations under randomised parameters, motivating falsifiable controls. Hewitt and Liang [18] introduce control tasks for language probing. Sparse Groundings uses nuisance-matched pools rather than random-label control tasks because the spatial-vision analogue of probe memorisation is patch-coordinate readout from the pooling rule.

**Vision-encoder representation analysis.** ViTs embed position via learned absolute, rotary, or relative-bias encodings and preserve patch-level localisation through depth [37]. Register tokens absorb high-norm patch artefacts [11]. CLIP carries a learned space preference [50]; effective receptive fields are visible in attention [26]. Park et al. [33] report that contrastive pretraining favours late-layer global / shape structure while masked pretraining favours early-layer local / texture structure — a precedent for the pretraining-objective $\rightarrow$ regime pattern we observe. Shtedritski et al. [39] use red-circle drawing as a controlled perturbation adjacent to our blank-canvas pool.

**Mechanistic interpretability for vision and VLMs.** SAEs originated for language model activations [5, 19] and have been extended to vision encoders [20, 21, 27, 32, 43, 53], including DINOv3 [43] and BioCLIP [41, 42]. SAE features can be non-localised relative to their causal drivers [16]. Templeton et al. [44] scale to 34M features in Claude 3 Sonnet; Basu et al. [3] extend causal tracing to MLLMs; Ameisen et al. [2] apply circuit-tracing to language models. Parodi et al. [34] show zero-ablation in DINO ViTs produces $\approx 250\times$ greater attention divergence than plausible replacements (mean / noise / shuffle); we adopt the same plausible-replacement framework for M4–M5 and the blank-canvas + matched-background pools as its pooling-protocol analogue.

| Representation | Module | object | blank | matched-bg | $\Delta_{\text{loc}}$ | Regime |
|---|---|---|---|---|---|---|
| *Base representations (raw vision-encoder patch features)* | | | | | | |
| CLIP ViT-B/16 | W3 colour | 0.989 | 0.125 | 0.970 | +0.018 | image-wide |
| | R2 scale-v2 | 0.581 | 0.250 | 0.613 | −0.031 | image-wide |
| | H1 relations | 0.981 | 0.775 | 0.931 | +0.050 | image-wide |
| | H2 counting | 0.932 | 0.255 | 0.675 | +0.256 | object-local |
| SigLIP 2 ViT-B/16 | W3 colour | 0.980 | 0.125 | 0.973 | +0.007 | image-wide |
| | R2 scale-v2 | 0.619 | 0.250 | 0.612 | +0.006 | image-wide |
| | H1 relations | 0.981 | 0.769 | 0.944 | +0.037 | image-wide |
| | H2 counting | 0.898 | 0.246 | 0.607 | +0.291 | object-local |
| BioCLIP ViT-B/16 | W3 colour | 1.000 | 0.125 | 1.000 | 0.000 | image-wide |
| | R2 scale-v2 | 0.556 | 0.250 | 0.512 | +0.044 | image-wide |
| | H1 relations | 0.969 | 0.781 | 0.938 | +0.031 | image-wide |
| | H2 counting | 0.790 | 0.207 | 0.711 | +0.079 | object-local |
| DINOv2-reg ViT-B/14 | W3 colour | 0.860 | 0.125 | 0.915 | −0.054 | image-wide |
| | R2 scale-v2 | 0.581 | 0.250 | 0.619 | −0.037 | image-wide |
| | H1 relations | 0.969 | 0.781 | 0.906 | +0.062 | object-local |
| | H2 counting | 0.430 | 0.394 | 0.510 | −0.080 | patch-position |
| DINOv2 no-reg ViT-B/14 | W3 colour | 0.797 | 0.125 | 0.871 | −0.074 | image-wide |
| | R2 scale-v2 | 0.575 | 0.250 | 0.613 | −0.038 | image-wide |
| | H1 relations | 0.950 | 0.744 | 0.900 | +0.050 | object-local |
| | H2 counting | 0.624 | 0.577 | 0.581 | +0.042 | patch-position |
| DINOv3 ViT-S/16 | W3 colour | 0.810 | 0.125 | 0.850 | −0.040 | image-wide |
| | R2 scale-v2 | 0.556 | 0.250 | 0.550 | +0.006 | image-wide |
| | H1 relations | 0.994 | 0.787 | 0.938 | +0.056 | object-local |
| | H2 counting | 0.964 | 0.233 | 0.836 | +0.128 | object-local |
| MAE ViT-B/16 | W3 colour | 0.998 | 0.125 | 0.997 | +0.001 | image-wide |
| | R2 scale-v2 | 0.644 | 0.250 | 0.594 | +0.050 | object-local |
| | H1 relations | 0.956 | 0.756 | 0.950 | +0.006 | image-wide |
| | H2 counting | 0.952 | 0.247 | 0.943 | +0.009 | image-wide |
| DeiT-B/16 distilled | W3 colour | 0.992 | 0.125 | 0.993 | −0.001 | image-wide |
| | R2 scale-v2 | 0.600 | 0.250 | 0.581 | +0.019 | image-wide |
| | H1 relations | 0.988 | 0.787 | 0.931 | +0.056 | object-local |
| | H2 counting | 0.817 | 0.289 | 0.759 | +0.058 | object-local |
| *SAE bases (sparse-autoencoder bases over the matching base representation)* | | | | | | |
| saev CLIP-B/16 24K | W3 colour | 0.973 | — | 0.596 | +0.378 | incomplete |
| | R2 scale-v2 | 0.544 | — | 0.413 | +0.131 | incomplete |
| | H1 relations | 0.975 | — | 0.756 | +0.219 | incomplete |
| | H2 counting | 0.816 | — | 0.561 | +0.255 | incomplete |
| saev DINOv2-reg 24K | W3 colour | 0.843 | — | 0.880 | −0.037 | incomplete |
| | R2 scale-v2 | 0.594 | — | 0.631 | −0.037 | incomplete |
| | H1 relations | 0.975 | — | 0.875 | +0.100 | incomplete |
| | H2 counting | 0.436 | — | 0.589 | −0.153 | incomplete |
| saev DINOv3-S/16 24K | W3 colour | 0.751 | 0.125 | 0.789 | −0.039 | image-wide |
| | R2 scale-v2 | 0.562 | 0.250 | 0.569 | −0.006 | image-wide |
| | H1 relations | 0.988 | 0.787 | 0.906 | +0.081 | object-local |
| | H2 counting | 0.948 | 0.232 | 0.751 | +0.196 | object-local |
| saev BioCLIP-B/16 24K | W3 colour | 0.996 | 0.125 | 0.946 | +0.050 | object-local |
| | R2 scale-v2 | 0.500 | 0.250 | 0.444 | +0.056 | object-local |
| | H1 relations | 0.906 | 0.731 | 0.613 | +0.294 | object-local |
| | H2 counting | 0.784 | 0.202 | 0.444 | +0.340 | object-local |

Table 11: Extended v2 module panel: 48 controlled-readout records over twelve audited representations $\times$ four v2 modules. $\Delta_{\text{loc}} = \text{object} - \text{matched-background}$; $\varepsilon = 0.05$. `incomplete` marks older SAE rows whose blank pool used the un-persisted preprocessing path; we report $\Delta_{\text{loc}}$ for those rows but not the regime. saev DINOv3-SAE matches CLIP-SAE's sharpens-locality pattern (object-local on H1 / H2 v2; image-wide on W3 / R2); saev BioCLIP-SAE is object-local on all four v2 modules.

| Representation | Pretraining family | Why in panel | Coverage |
|---|---|---|---|
| *Base representations (raw vision-encoder patch features)* | | | |
| CLIP ViT-B/16 [36] | softmax-contrastive | default CLIP | R1 + v2 |
| SigLIP 2 ViT-B/16 [49] | sigmoid + dense pretraining | dense-localisation claim | R1 + v2 |
| DINOv2 ViT-B/14 + reg [11, 31] | SSL + registers | default DINOv2 | R1 + v2 |
| DINOv3 ViT-S/16 [40] | SSL + Gram anchoring | dense-feature claim | R1 + v2 |
| MAE ViT-B/16 [17] | masked reconstruction | non-discriminative SSL | R1 + v2 |
| DeiT-B/16 distilled [48] | supervised + distillation | supervised baseline | R1 + v2 |
| DINOv2 ViT-B/14 (no reg) [31] | SSL, no registers | register effect | v2 only |
| BioCLIP ViT-B/16 [41] | contrastive on TreeOfLife-10M | domain-specific base | v2 only |
| *SAE bases (sparse-autoencoder bases over the matching base)* | | | |
| saev CLIP-B/16 24K [43] | SAE over CLIP-B/16 | default vision SAE | R1 + v2 |
| saev DINOv2-reg 24K [43] | SAE over DINOv2-reg | SAE-basis preservation | R1 + v2 |
| saev DINOv3-S/16 24K [43] | SAE over DINOv3-S/16 | SAE over Gram-anchored SSL | v2 only |
| saev BioCLIP-B/16 24K [43] | SAE over BioCLIP | SAE over domain-specific base | v2 only |

Table 12: Audited-representation panel. "R1 + v2" marks representations with R1 position results plus v2 module results (W3, R2, H1, H2, H3) in this paper; "v2 only" marks the four extended-panel rows that have v2 module coverage but not R1. Base representations and SAE bases are evaluated separately; an SAE result is a property of the SAE basis, not of the underlying base encoder.

| Task | Selector | Seed gap range | LOO gap range | Most-sensitive drop |
|---|---|---|---|---|
| D1 ADE20K seg | `where_r1_locality` | $[+0.057, +0.514]$ | $[-0.300, +0.600]$ | MAE |
| D2 NYUv2 depth | `geometry_r2_h1_loc.` | $[+0.468, +0.818]$ | $[+0.616, +1.110]$ | DINOv2-reg |
| D4 iNat-mini cls | `what_w1_w3_loc.` | $[-0.571, -0.171]$ | $[-0.800, -0.400]$ | MAE |
| D5 FSC-147 count | `structure_h2_loc.` | $[-0.286, +0.514]$ | $[-0.200, +0.200]$ | CLIP |

Table 13: Two-dimensional sensitivity for the four main cached-real downstream receipts. Seed gap range is across the five head seeds at fixed encoder set. LOO gap range is across the six leave-one-encoder-out panels at fixed seed. The most-sensitive drop column names the encoder whose removal moves the selected gap furthest from the full-panel value. D2 is the only receipt that stays positive across both sensitivity dimensions; D1 is seed-stable but sign-flips under LOO when MAE is dropped, mirroring the D3 leave-one-out fragility reported in Appendix A.11.

| Task | Selector | $n = 6$ | $n = 8$ | $\Delta$ | Status |
|---|---|---|---|---|---|
| D1 ADE20K seg | `where_r1_loc.` | $+0.286$ | $+0.381$ | $+0.095$ | panel-robust + |
| D2 NYUv2 depth | `geometry_r2_h1_loc.` | $+0.818$ | $-0.210$ | $-1.028$ | panel-flip |
| D4 iNat-mini cls | `what_w1_w3_loc.` | $-0.571$ | $-0.132$ | $+0.439$ | sign-fragile $n = 8$ |
| D5 FSC-147 count | `structure_h2_loc.` | $0.000$ | $0.000$ | $0.000$ | sign-fragile |

Table 14: Validation rank-correlation under panel expansion from $n = 6$ to $n = 8$. Both panels use the same locked task-family selectors, five-seed median downstream summaries, and identical task definitions; only the encoder set changes (the $n = 8$ panel adds BioCLIP raw and DINOv2 no-reg raw). Selector completeness verified: $n_{\text{cells actual}} = 8$ for every selected task, no rows filtered. $\Delta$ is the $n = 8$ minus $n = 6$ selected $\rho$-gap. D1 stays positive across panels and strengthens; D2's strongest-positive result at $n = 6$ flips negative at $n = 8$; D4 attenuates substantially toward zero; D5 stays at the fragile null.

| Representation | Acc@0.5 | $\Delta$ vs area prior | Beats oracle |
|---|---|---|---|
| SigLIP 2-B/16 raw | 0.547 | $+0.051$ | yes |
| DINOv3-S/16 raw | 0.531 | $+0.035$ | yes |
| DeiT-B/16 raw | 0.527 | $+0.031$ | yes |
| DINOv2-reg B/14 raw | 0.512 | $+0.016$ | yes |
| CLIP-B/16 raw | 0.453 | $-0.043$ | no |
| MAE-B/16 raw | 0.305 | $-0.191$ | no |

Table 15: D3 RefCOCO+ candidate-region grounding, six-base panel. Acc@0.5 over 768 expressions; $\Delta$ vs area prior at Acc@0.5 $= 0.496$ (centre $0.320$, random $0.486$).

| | Object pool | | Blank same-fp pool | | Matched-bg pool | |
|---|---|---|---|---|---|---|
| Representation | Acc | Selectivity | Acc | Selectivity | Acc | Selectivity |
| CLIP-B/16 raw | 0.475 | 0.414 | 0.993 | 0.943 | 0.855 | 0.803 |
| SigLIP 2-B/16 raw | 0.991 | 0.941 | 0.998 | 0.948 | 0.581 | 0.523 |
| MAE-B/16 raw | 0.827 | 0.774 | 0.983 | 0.933 | 0.838 | 0.785 |
| DeiT-B/16 raw | 0.794 | 0.740 | 0.998 | 0.948 | 0.553 | 0.494 |
| DINOv2-reg B/14 raw | 0.638 | 0.581 | 0.999 | 0.949 | 0.705 | 0.650 |
| DINOv3-S/16 raw | 0.822 | 0.769 | 0.999 | 0.949 | 0.789 | 0.735 |

Table 16: R1 49-way position-probe selectivity across the six-base panel and three pools. "Acc" is balanced accuracy on the held-out 20% object split with the locked logistic probe; "Selectivity" is the chance-normalised true score minus the larger of the chance-normalised five-seed shuffled-label median and the chance-normalised footprint-count metadata-oracle score. Five-seed shuffled-label probes converge near chance ($1/49 \approx 0.020$; per-row range $0.004$–$0.037$), and the metadata oracle using footprint and matched-bg counts at $p \in \{14, 16\}$ reaches $0.069$. All 18 (representation, pool) rows verdict `selective`.

| Representation | object | blank | matched-bg | $\Delta_{\text{inc}}$ | $\Delta_{\text{loc}}$ | regime |
|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.989 | 0.125 | 0.970 | +0.864 | +0.018 | image-wide |
| DINOv2-reg raw | 0.860 | 0.125 | 0.915 | +0.735 | −0.054 | image-wide |
| SigLIP 2-B/16 raw | 0.980 | 0.125 | 0.973 | +0.855 | +0.007 | image-wide |
| DINOv3-S/16 raw | 0.810 | 0.125 | 0.850 | +0.685 | −0.040 | image-wide |
| MAE-B/16 raw | 0.998 | 0.125 | 0.997 | +0.873 | +0.001 | image-wide |
| DeiT-B/16 raw | 0.992 | 0.125 | 0.993 | +0.867 | −0.001 | image-wide |

Table 17: W3 colour attribute under three pools. All six base representations are image-wide on colour: blank pool exactly at chance, matched-background pool reaches the same accuracy as the object pool ($|\Delta_{\text{loc}}| < \varepsilon = 0.05$), incremental delta large. Object-pool ceilings split into a high tier (MAE, DeiT, CLIP, SigLIP 2 $\geq 0.98$) and a low tier (DINOv2-reg 0.860, DINOv3 0.810).

| Representation | object | blank | matched-bg | patch-count | $\Delta_{\text{inc}}$ | $\Delta_{\text{loc}}$ | regime |
|---|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.581 | 0.250 | 0.613 | 0.250 | +0.331 | −0.031 | image-wide |
| DINOv2-reg raw | 0.581 | 0.250 | 0.619 | 0.250 | +0.331 | −0.037 | image-wide |
| SigLIP 2-B/16 raw | 0.619 | 0.250 | 0.612 | 0.250 | +0.369 | +0.006 | image-wide |
| DINOv3-S/16 raw | 0.556 | 0.250 | 0.550 | 0.250 | +0.306 | +0.006 | image-wide |
| MAE-B/16 raw | 0.644 | 0.250 | 0.594 | 0.250 | +0.394 | +0.050 | object-local |
| DeiT-B/16 raw | 0.600 | 0.250 | 0.581 | 0.250 | +0.350 | +0.019 | image-wide |

Table 18: R2 scale-v2 (footprint-fixed; centre-position, four zooms). Patch-count baseline at chance (0.250) for every representation; every representation clears the $\Delta_{\text{inc}} \geq 0.20$ DoD; range $+0.306$ (DINOv3) to $+0.394$ (MAE). MAE is the lone object-local representation ($\Delta_{\text{loc}} = +0.050$, just over $\varepsilon$); the five remaining representations are image-wide with $|\Delta_{\text{loc}}| \leq 0.04$.

| Representation | object | blank | matched-bg | indep-pos | $\Delta_{\text{inc}}$ / $\Delta_{\text{loc}}$ | regime |
|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.981 | 0.775 | 0.931 | 0.869 | +0.206 / +0.050 | object-local |
| DINOv2-reg raw | 0.969 | 0.781 | 0.906 | 0.869 | +0.188 / +0.062 | object-local |
| SigLIP 2-B/16 raw | 0.981 | 0.769 | 0.944 | 0.869 | +0.212 / +0.037 | image-wide |
| DINOv3-S/16 raw | 0.994 | 0.787 | 0.938 | 0.869 | +0.206 / +0.056 | object-local |
| MAE-B/16 raw | 0.956 | 0.756 | 0.950 | 0.869 | +0.200 / +0.006 | image-wide |
| DeiT-B/16 raw | 0.988 | 0.787 | 0.931 | 0.869 | +0.200 / +0.056 | object-local |

Table 19: H1 relations under three pools + independent-position baseline. All six base representations beat the independent-position baseline by $+0.088$ (MAE) to $+0.125$ (DINOv3). Two encoders are object-local on H1 ($\Delta_{\text{loc}} \geq \varepsilon + \delta$): CLIP and DINOv2-reg. Two are borderline object-local ($\varepsilon \leq \Delta_{\text{loc}} < \varepsilon + \delta$): DeiT and DINOv3, both at $\Delta_{\text{loc}} \approx +0.056$. SigLIP 2 and MAE are image-wide.

| Representation | object | blank | matched-bg | fp-area | $\Delta_{\text{loc}}$ | $\Delta_{\text{inc-dual}}$ |
|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.932 | 0.255 | 0.675 | 0.382 | $+0.256$ | $+0.550$ |
| DINOv2-reg raw | 0.430 | 0.394 | 0.510 | 0.382 | $-0.080$ | $+0.036$ |
| SigLIP 2-B/16 raw | 0.898 | 0.246 | 0.607 | 0.382 | $+0.291$ | $+0.516$ |
| DINOv3-S/16 raw | 0.964 | 0.233 | 0.836 | 0.382 | $+0.128$ | $+0.583$ |
| MAE-B/16 raw | 0.952 | 0.247 | 0.943 | 0.382 | $+0.009$ | $+0.571$ |
| DeiT-B/16 raw | 0.817 | 0.289 | 0.759 | 0.382 | $+0.058$ | $+0.435$ |

Table 20: H2 v2 counting (target-scene-area $T = 2304$ px). Footprint-area baseline at 0.382 (vs 1.000 in v1, Appendix Table 9). Five of six representations clear the $+0.15$ dual-nuisance DoD. CLIP and SigLIP 2 are object-local on H2 v2 ($\Delta_{\text{loc}} \geq 0.05$); DINOv3 and DeiT are weakly object-local; MAE is image-wide; DINOv2-reg is the lone fail.

| Representation | object | blank | matched-bg | distractor | metadata | target$-$distr. | regime |
|---|---|---|---|---|---|---|---|
| CLIP-B/16 raw | 0.953 | 0.250 | 0.531 | 0.391 | 0.188 | $+0.562$ | object-local |
| DINOv2-reg raw | 0.625 | 0.250 | 0.516 | 0.281 | 0.188 | $+0.344$ | object-local |
| SigLIP 2-B/16 raw | 0.781 | 0.250 | 0.672 | 0.391 | 0.188 | $+0.391$ | object-local |
| DINOv3-S/16 raw | 0.766 | 0.250 | 0.422 | 0.281 | 0.188 | $+0.484$ | object-local |
| MAE-B/16 raw | 0.688 | 0.250 | 0.797 | 0.516 | 0.188 | $+0.172$ | image-wide |
| DeiT-B/16 raw | 0.969 | 0.250 | 0.656 | 0.359 | 0.188 | $+0.609$ | object-local |

Table 21: H3 distractor robustness. The probe decodes target-object colour in scenes with 0, 1, 3, or 5 distractors. Metadata-only and blank controls are below the object pool for all six base representations. Five of six are object-local by target-vs-distractor locality; MAE is image-wide because matched-background features exceed the target-object pool.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state the three contributions (audit protocol, empirical audit across six raw encoders and two SAE bases, released library) and list every numerical claim with its scope (six- vs eight-encoder panel, locked selectors).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §7 discusses the synthetic-stimuli scope, $n = 6$ bootstrap-CI width, panel-conditioned downstream-rank correlations, and unevaluated robustness perturbations (continuous jitter, off-grid centres, alternative pooling, blank intensity / noise / texture, footprint jitter, cutout).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3 formalises the three pools, $\Delta_{\text{inc}}$, $\Delta_{\text{loc}}$, and the regime threshold $\varepsilon$. Every table and figure is backed by a JSON receipt path in the released artifact. The released library and stimulus generators (Appendix A, §A.9) reproduce all main-text numbers from cached features; the matched-background pool uses a deterministic `blake2b` digest for byte-reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The repository ships a `pip`-installable library, CLI, stimulus generators, evaluation cards, locked selector registry, and a cached-feature reproduction manifest. `docs/REPRODUCING.md` documents installation, smoke test, and the per-figure / per-table workflow. Anonymous access is provided via the submission page.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Probes use scikit-learn linear models with documented $\alpha$, train / test splits (80 / 20), and seeds. Encoder layer choices, pooling, register-stripping, manifest checksums, and selector-registry locks are stated in §3, §5, and Appendix A.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The validation results (§6, Table 2, Appendix A §A.10) are reported as five-seed median Spearman $\rho$-gaps with seed-range and leave-one-out ranges, plus an $n = 8$ panel-expansion sensitivity. The threshold $\varepsilon = 0.05$ is set to the seed-stability scale measured under the controls. Bootstrap CIs at $n = 6$ are noted as wide and treated as a sensitivity layer rather than a power claim.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: §7 "Compute footprint" reports total compute to produce the submitted empirical receipts ($\sim$ 25–40 GPU-hours on two NVIDIA RTX 4090s, plus $\sim$ 20–40 CPU-hours for probe fitting, aggregation, and figure / table generation), the cached-feature disk footprint ($\sim$ 3 GB), and the cached-reproduction cost (no GPU forwards; a few CPU-minutes).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The work uses publicly released vision encoders, sparse-autoencoder bases, and standard public datasets, with no human subjects, no scraped private data, and no deployment claims.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Positive impacts: Sparse Groundings tightens the falsifiability of vision-encoder representation claims by making each claim testable against named pooling controls, which improves transparency around what frozen vision encoders actually represent and reduces the risk that downstream stacks build on confounded probe results. Potential negative impacts and misuse: "object-local" or "image-wide" regime labels should not be read as safety, fairness, perceptual-competence, or deployment-readiness certifications; the natural-image bridge is a transfer check, not a generalisation guarantee; and the synthetic stimulus suite is a controlled diagnostic, so rankings produced under it should not be marketed as evidence of overall vision-encoder quality. We discuss intended use and non-use in §7.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artefact is a benchmark protocol and analysis library; it does not release new model weights or scraped datasets.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All vision encoders, sparse-autoencoder bases, and datasets are cited in §2 and Appendix A (§A.13). Each asset is used under its original license; the released artefact carries an Apache-2.0 license.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark, library, evaluation cards, and selector registry are documented in the repository (`README.md`, `docs/REPRODUCING.md`, eval cards under `paper/eval-cards/`) and in Appendix A §A.9.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core methodology. Any LLM use was for writing assistance and does not impact the empirical results.