# WHAT POOLED EXPLAINED VARIANCE CAN HIDE: EVALUATING CROSSCODER FIDELITY IN VISION ENCODERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Crosscoders, a variant of sparse autoencoders, were initially developed to compare representations across models and to analyze features across layers. However, their use for comparing blocks within a single vision encoder remains underexplored. We ask whether raw explained variance reflects typical-token reconstruction and predicts VLM behavior after a reconstruction is placed back into the encoder. We study this limitation using crosscoders, which reconstruct representations across layers and models. Our analysis shows that pooled explained variance weights each token by its squared distance from the activation mean, allowing a small high-norm population to dominate the score and obscure weaker reconstruction elsewhere. We evaluate crosscoders across five vision encoders using masked, token-balanced, and norm-stratified fidelity, separately accounting for population recentering. Controlled sparsity sweeps reveal that typical-token reconstruction deteriorates faster than raw explained variance as crosscoders become sparser. We then test these reconstructions inside four VLMs on the SpatiaLab spatial-reasoning benchmark. In Qwen2.5-VL, reconstructions explain over 99% of pooled activation variance yet reduce accuracy by up to 6.9 percentage points. Additional encoders and a separate stress test support the broader discrepancy between pooled fidelity and preserved behavior. Interventions identify model-specific multilayer perceptron contributions necessary for the measured high-norm contrasts, which later blocks often inherit, while Gemma provides a control without a persistent high-norm tail. Together, these findings establish the need to evaluate reconstruction across token populations and test its behavioral consequences. We recommend reporting raw explained variance alongside token-balanced and norm-stratified fidelity, followed by block replacement to assess whether reconstruction errors affect the behavior being studied.

## 1 INTRODUCTION

Near-perfect reconstruction scores do not establish that a sparse representation preserves a model's behavior. Sparse autoencoders express dense activations through learned features; crosscoders extend this approach across layers and models to study how features persist and change (Gao et al., 2024; Lindsey et al., 2024). Increasing sparsity simplifies feature analysis but introduces reconstruction errors. Before interpreting these features, we therefore need to establish which parts of a representation they preserve. Prior evaluations of sparse representations motivate testing reconstruction fidelity alongside downstream behavior (Gao et al., 2024; Karvonen et al., 2025; Chatzoudis et al., 2026). We examine a specific obstacle to this assessment: pooled explained variance gives visual tokens unequal influence over the reconstruction score. For simplicity, we refer to explained variance as EV henceforth.

Several vision encoders develop a small population of unusually high-norm tokens (Darcet et al., 2024). Some carry high-level semantic information and support visual reasoning (Luo et al., 2026). Pooled EV weights each token by its squared distance from the activation mean. When a small population accounts for most of the variance, accurately reconstructing it dominates the score and obscures weaker reconstruction elsewhere. Preserving these high-norm tokens is important, but it is not sufficient to establish the fidelity of the remaining visual representation. This raises our central



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

question: does high pooled explained variance describe reconstruction across visual tokens, and does that reconstruction preserve the VLM's behavior?

We address this question across five vision encoders. **Our contributions are twofold.** First, we separate pooled explained variance token weighting from population recentering and identify model-specific MLP contributions necessary for the measured norm contrasts. Second, sparsity sweeps show that typical-token fidelity deteriorates faster than raw EV, while block replacements in four VLMs on SpatiaLab show that similar pooled scores accompany different model behavior. These findings motivate reporting pooled EV alongside token-balanced and norm-stratified fidelity and testing reconstructions inside the original model.

## 2 What pooled explained variance measures

### 2.1 Crosscoder reconstruction

A crosscoder reads a token's residual activations from all vision blocks, maps them to one sparse code, and reconstructs each block through a separate decoder. Training combines squared reconstruction error with a decoder-norm-weighted $\ell_1$ penalty (Appendix Equation 5). We report realized $\ell_0$, the average number of active features per token, to measure sparsity.

For a fixed vision block, let $x_i$ denote the clean activation of token $i$ and $\hat{x}_i$ its reconstruction. Let $S$ contain the evaluated tokens and $\mu = |S|^{-1} \sum_{i \in S} x_i$ be their mean clean activation. We define the squared reconstruction error as $e_i = \|x_i - \hat{x}_i\|^2$ and the squared distance from the mean as $v_i = \|x_i - \mu\|^2$. These quantities separate the error in reconstructing a token from its contribution to the total activation variance.

### 2.2 Which tokens determine pooled EV?

Pooled EV compares total reconstruction error with total squared distance from the activation mean. When every $v_i > 0$, it also has the following weighted-average form:

$$\mathrm{EV}_{\mathrm{raw}} = 1 - \frac{\sum_{i \in S} e_i}{\sum_{i \in S} v_i} = \sum_{i \in S} \underbrace{\frac{v_i}{\sum_{j \in S} v_j}}_{\text{token weight}} \left(1 - \frac{e_i}{v_i}\right). \tag{1}$$

Each token receives a weight proportional to its squared distance from the mean. Consequently, accurately reconstructing tokens that account for most of the variance dominates the pooled score, even when reconstruction is weaker elsewhere. The relevant quantity is distance from the mean, rather than activation norm alone.

To measure this concentration, we sort tokens by their clean activation norm and divide them into ten equal-count groups, $D_1, \ldots, D_{10}$. Their contributions satisfy

$$\begin{aligned} \alpha_d &= \frac{\sum_{i \in D_d} v_i}{\sum_{i \in S} v_i}, \qquad \mathrm{EV}_d^{(g)} = 1 - \frac{\sum_{i \in D_d} e_i}{\sum_{i \in D_d} v_i}, \\ \mathrm{EV}_{\mathrm{raw}} &= \sum_{d=1}^{10} \alpha_d \, \mathrm{EV}_d^{(g)}. \end{aligned} \tag{2}$$

Here, $g$ indicates that every group uses the same block-wide mean. The weight $\alpha_{10}$ shows how strongly the highest-norm decile influences pooled EV. Equal contributions across the ten deciles give $\alpha_{10} = 0.1$. This weight depends only on clean activations, not on reconstruction quality.

### 2.3 Evaluating reconstruction across token populations

The decomposition explains why a pooled score alone does not describe reconstruction across tokens. We therefore use three complementary summaries, each answering a different question.

**Masked EV.** For the common threshold-based audit, we exclude tokens whose clean activation norm exceeds ten times the block median at any vision block. This union mask selects the same

retained population across blocks. For a fixed-size comparison, D10-masked EV instead excludes the highest-norm decile separately at each block. Both scores recompute the mean on the retained tokens. Appendix D.2 gives the model-specific sweep masks and robustness checks.

**Token-balanced EV.** We average each token's normalized error while retaining the block-wide mean:

$$\mathrm{EV}_{\mathrm{tok}} = 1 - \frac{1}{|S|} \sum_{i \in S} \frac{e_i}{\max(v_i, \epsilon)}. \tag{3}$$

The positive numerical floor $\epsilon$ prevents division by zero and limits the influence of denominators close to zero. Unlike pooled EV, this score assigns equal weight to each token's normalized error.

**Norm-decile EV.** At each block, we compute EV separately within each clean-activation norm decile, using that decile's own mean. Norm-decile EV averages all ten scores. We call the average over deciles 1–9 *typical-token EV* and report decile 10 separately. This gives *typical* an explicit definition and reveals differences across norm ranges without selecting a magnitude threshold.

Changing the mean also changes the comparison. For a fixed group $D$, let $\mu_D$ be its mean, $S_D = \sum_{i \in D} e_i$ its reconstruction error, $W_D = \sum_{i \in D} \|x_i - \mu_D\|^2$ its within-group variation, and $B_D = |D| \|\mu_D - \mu\|^2$ the contribution from its displacement from the block-wide mean. The scores using the block-wide and group means, respectively, satisfy, for $W_D > 0$,

$$\begin{aligned} Q_D^{(g)} &= 1 - \frac{S_D}{W_D + B_D}, \qquad Q_D^{(\ell)} = 1 - \frac{S_D}{W_D}, \\ Q_D^{(g)} - Q_D^{(\ell)} &= \frac{S_D B_D}{W_D (W_D + B_D)} \geq 0. \end{aligned} \tag{4}$$

Thus, recentering reduces or leaves unchanged the score even when the evaluated tokens and reconstruction errors remain identical. We measure this effect separately, so differences between pooled, masked, and typical-token EV are not attributed entirely to unequal token weights.

## 3 Experimental setup

We evaluate whether reconstruction fidelity across visual tokens translates into preserved VLM behavior, then test which components produce the high-norm contrasts.

**Models and crosscoder training.** We study the vision encoders of Qwen2.5-VL-3B-Instruct (Bai et al., 2025), LLaVA-OneVision-1.5-4B (An et al., 2025), InternVL2.5-4B (Chen et al., 2024), Phi-3.5-Vision (Abdin et al., 2024), and Gemma-3-4B SigLIP (Gemma Team, 2025). We collect residual outputs using 600 images from CV-Bench (Tong et al., 2024), with training and validation split by image (Appendix B). Downstream evaluation images are separate from crosscoder training. We audit reconstruction across all five encoders and run controlled sparsity sweeps in Qwen, LLaVA, InternVL, and Phi. Gemma uses a single dictionary (Appendix F). The final sweep settings use 20 training epochs and one initialization seed each. Appendix D.5 lists dictionary widths and sparsity settings for the sweeps; Appendix B gives the training and implementation details.

**Reconstruction and behavioral evaluation.** We report the reconstruction scores defined in Section 2. For behavioral evaluation, we use all 1,400 SpatiaLab multiple-choice questions spanning six spatial tasks (Wasi et al., 2026). We evaluate six frozen dictionaries per model in Qwen, Phi, LLaVA, and InternVL, replacing one middle or late vision-block output at a time. Each reconstruction uses the full clean activation stack, including later blocks, and replaces one block's output before the remaining model runs normally. This evaluates offline reconstruction replacement; it does not test a substitute computed solely from the current and earlier blocks. We measure accuracy, agreement with the unreplaced model, correct-answer margin change, and KL divergence between clean and reconstructed answer-choice distributions. Accuracy comparisons use image-level paired sign-flip tests, paired bootstrap 95% confidence intervals, and Holm correction within each model. Per-task results and correlations are descriptive. Appendix B.3 gives the evaluated blocks and evaluation protocol; Appendix B.4 details the paired tests, confidence intervals, and multiple-comparison corrections.

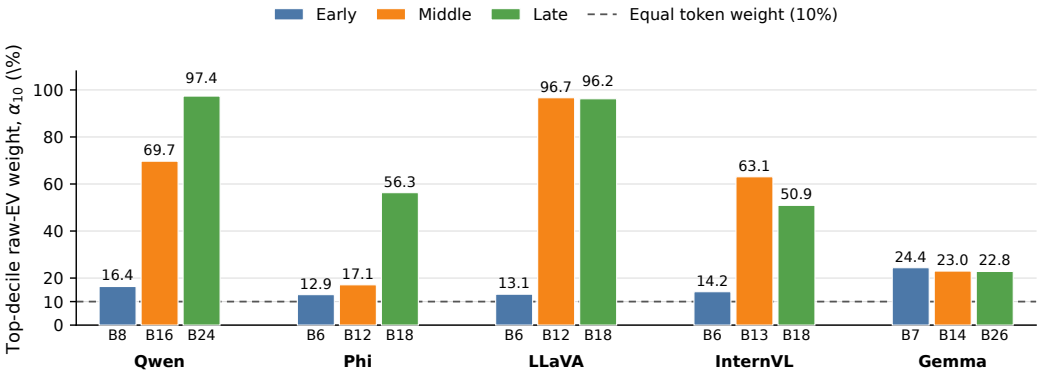



Figure 1: **Empirical test of the raw-EV weighting term.** Each bar shows $\alpha_{10}$ from Equation 2: the fraction of the total squared distance from the activation mean contributed by norm decile 10. Every decile contains 10% of the tokens, so the dashed line is the equal-weight reference. Values use all 1,400 SpatiaLab images.

**Token and component interventions.** A separate stress test evaluates the four VLMs on selected Visual Genome images (Krishna et al., 2017) (Appendix A.1). On these images, norm-decile restoration and replacement test which token groups affect behavior (Appendices A and A.4). For the mechanism analysis, all five encoders use the same 100 SpatiaLab images sampled uniformly without replacement with seed 2027. We replace selected tokens' attention or MLP contributions with the ordinary-token mean, then run later blocks normally. We repeat the intervention at intermediate MLP stages to test which contributions are necessary for the measured norm contrast. Gemma serves as a control for this analysis. Appendix E specifies the frozen sample, token-selection rules, intervention sites, and controls; Table A18 reports the propagated effects near the encoder's end.

## 4 Results across vision encoders

### 4.1 A small token population dominates pooled variance

Figure 1 shows that equally sized token groups make sharply unequal contributions to pooled variance. The highest-norm decile dominates pooled variance at the late evaluated blocks in Qwen and LLaVA, while Phi and InternVL show weaker concentration. Through Equation 2, this establishes which token population receives most of the weight in raw EV. A high pooled score therefore provides limited evidence about reconstruction outside that population. Gemma lacks a comparable persistent tail (Appendix F), showing that the severity of this imbalance depends on the encoder.

Figure 2 reveals the consequence for reconstruction assessment: near-perfect pooled EV coexists with substantially weaker fidelity after the highest-norm tokens are excluded. In Qwen, token-balanced EV also remains high where masked scores are low. Giving tokens equal weight therefore does not answer the same question as measuring variation within a retained population. The complementary scores in Section 2 expose distinctions that a single reconstruction summary conceals.

The final-block jumps in Qwen ($31 \rightarrow 32$) and Gemma ($26 \rightarrow 27$) occur because centered activation variance grows faster than squared reconstruction error, even though absolute error increases. Phi's raw-EV rise at block $12 \rightarrow 13$ follows the same pattern; its subsequent decline through block 23 reflects faster error growth than variance growth. Appendix Table A12 reports the per-token error and variance ratios underlying these changes. These trends describe crosscoder fidelity, not changes in encoder capability.

The fixed-population comparison in Appendix D.4 explains why these differences cannot be attributed entirely to unequal weighting. For the same retained Qwen tokens and reconstruction errors, centering on the retained population produces a substantially lower score than centering on the full population. The effect is smaller in the other encoders. As Equation 4 shows, changing the mean changes the

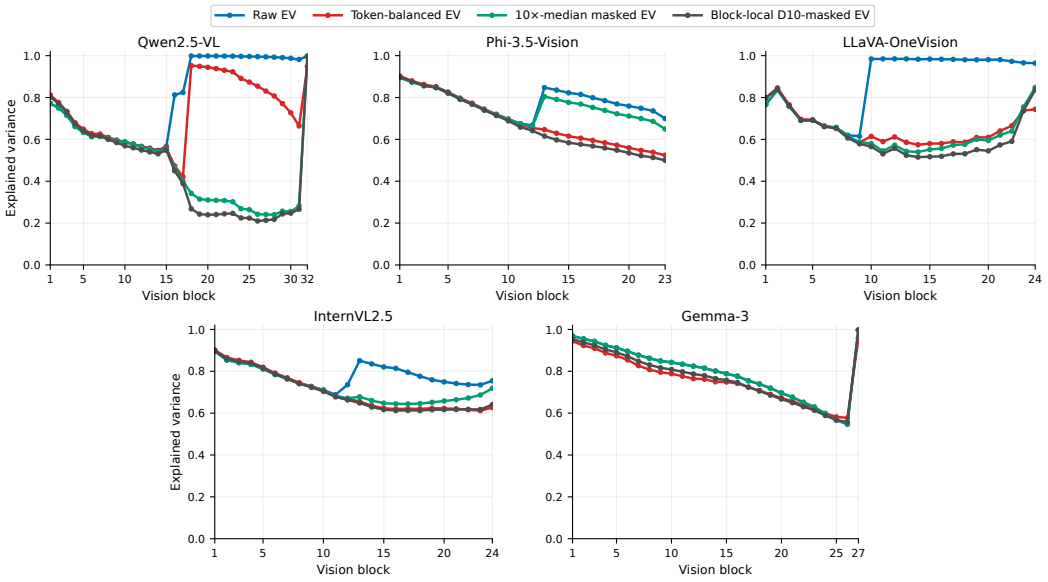



Figure 2: **Full-depth fidelity on SpatiaLab.** Raw EV (blue), token-balanced EV (red), 10×-median union-masked EV (green), and block-local D10-masked EV (dark gray) use all 1,400 images and one representative crosscoder per model. The fixed-threshold rule removes a token at every depth if it crosses the threshold at any block; D10 masking independently removes the highest-norm 10% at each block. Both masked scores recenter the retained population.

Table 1: Sweep endpoints at fixed dictionary width 4096 on the CV-Bench Depth validation split, from densest to sparsest. The sparsity penalty $\lambda$ increases from .1 to 10 for Phi, Qwen, and InternVL, and from .1 to 3 for LLaVA. Raw EV pools reconstruction error and block-centered variance across tokens and blocks after the training normalization. "Typ." averages the corresponding pooled scores for norm deciles 1–9, with deciles defined separately at each block. Gaps (Raw minus Typ.) are computed before rounding.

| Encoder | $\ell_0$ | Raw | Typ. | Gap |
|---|---|---|---|---|
| Phi | 2241→26 | .959→.705 | .922→.459 | .037→.246 |
| Qwen | 2339→58 | .996→.975 | .865→.236 | .130→.739 |
| LLaVA | 2404→76 | .991→.973 | .878→.620 | .114→.353 |
| Intern | 2611→23 | .945→.677 | .904→.446 | .041→.231 |

reference variation without changing the reconstruction. Interpreting a fidelity gap therefore requires specifying both the token population and the mean used to evaluate it.

### 4.2 Sparsity widens the reconstruction gap

Table 1 holds dictionary width fixed at 4096 and shows that the discrepancy grows as the sparsity penalty increases and fewer features remain active. Across these fixed-width sweeps, typical-token EV deteriorates faster than raw EV as realized $\ell_0$ falls. The contrast is strongest in Qwen, where pooled fidelity changes little despite a pronounced loss of typical-token fidelity. These scores summarize reconstruction across blocks on the CV-Bench Depth validation split; Appendix D.5 gives the full settings. The result directly affects the evaluation of sparse representations: the apparent reconstruction cost of using fewer active features depends strongly on which token populations the score emphasizes.

Raw and masked EV retain the same within-block dictionary ordering in these sweeps. Their agreement on ordering does not make their magnitudes interchangeable. Pooled EV communicates a much smaller reconstruction loss under increasing sparsity, so reporting it alone obscures how much fidelity is lost outside the highest-norm group.

Table 2: Overall SpatiaLab accuracy (%). Each panel reports six frozen dictionaries at two replacement blocks $B$; the clean accuracy appears beside the model name. All models use all 1,400 images, and $d$ is dictionary width. † marks an accuracy loss relative to the clean model that is statistically significant after Holm correction across the 12 comparisons within that model ($p < .05$). Full per-task results are in Appendix C.1. Bold indicates the highest replacement accuracy at each block.

| **Qwen2.5-VL-3B-Instruct** (clean: 48.4) | | | | **Phi-3.5-Vision-4B-Instruct** (clean: 46.1) | | | |
|---|---|---|---|---|---|---|---|
| **Dictionary** | | **Accuracy** | | **Dictionary** | | **Accuracy** | |
| $\lambda$ | $d$ | $B = 16$ | $B = 24$ | $\lambda$ | $d$ | $B = 12$ | $B = 18$ |
| .1 | 2048 | **48.0** | 48.1 | .1 | 4096 | 45.9 | 46.2 |
| .1 | 4096 | 47.9 | **48.3** | .1 | 8192 | **46.1** | **46.4** |
| 1 | 2048 | 45.7 | 43.1† | .3 | 4096 | 45.7 | 46.2 |
| 1 | 4096 | 45.6 | 43.0† | 1 | 4096 | 44.4 | 44.5 |
| 10 | 2048 | 43.3† | 41.5† | 3 | 4096 | 42.7† | 40.4† |
| 10 | 4096 | 43.3† | 42.8† | 10 | 4096 | 40.1† | 38.2† |
| **LLaVA-OneVision-1.5-4B-Instruct** (clean: 52.9) | | | | **InternVL2.5-4B** (clean: 50.0) | | | |
| **Dictionary** | | **Accuracy** | | **Dictionary** | | **Accuracy** | |
| $\lambda$ | $d$ | $B = 12$ | $B = 18$ | $\lambda$ | $d$ | $B = 13$ | $B = 18$ |
| .1 | 2048 | 52.8 | 51.9 | .1 | 2048 | 50.3 | 49.6 |
| .1 | 4096 | **52.9** | **52.9** | .1 | 4096 | **50.4** | 49.5 |
| 1 | 2048 | 49.7† | 49.5† | 1 | 2048 | 48.2 | 48.4 |
| 1 | 4096 | 51.5 | 50.7 | 1 | 4096 | 49.0 | **49.7** |
| 3 | 2048 | 47.1† | 47.8† | 10 | 2048 | 43.9† | 44.8† |
| 3 | 4096 | 46.5† | 49.3† | 10 | 4096 | 43.6† | 46.0† |

The discrepancy also occurs in single-block BatchTopK sparse autoencoders (Bussmann et al., 2024) trained without an $\ell_1$ penalty. On Qwen block 24, pooled EV remains near one while typical-token fidelity declines as sparsity increases (Appendix D.6, Table A15). Thus, the discrepancy is not restricted to $\ell_1$-regularized crosscoders.

### 4.3 High pooled EV does not establish preserved behavior

The SpatiaLab block replacements in Table 2 establish that reconstruction errors also have behavioral consequences. Dense reconstructions achieve accuracy close to the clean baselines, while sparse replacements produce losses that remain significant after Holm correction in every evaluated VLM. Qwen provides the clearest mismatch: across six block-24 reconstructions evaluated on the same 1,400 SpatiaLab images, raw EV exceeds .995 while accuracy drops by up to 6.9 percentage points.

LLaVA shows a similar mismatch between a narrow raw-EV range and changing behavior. Phi and InternVL reproduce the accuracy loss under increasing sparsity, but their raw EV also falls more substantially. The evidence therefore establishes a shared sparsity pattern with different degrees of pooled-score compression across encoders.

At fixed dictionary width and the late evaluated block, increasing sparsity shifts the per-image clean-to-reconstruction KL distributions toward larger values in all four VLMs, indicating greater changes in answer-choice probabilities (Appendix C.2). Per-task losses vary across spatial categories; those comparisons are descriptive (Appendix C.1).

### 4.4 Tokens outside the highest-norm group affect behavior

The stress test on a separate evaluation set (Appendix A.1) supports the same qualitative discrepancy (Appendix A). For Qwen's sparsest block-16 reconstruction, restoring decile 8 yields a larger observed accuracy recovery than restoring the highest-norm decile. The decile-8 gain over full reconstruction remains significant after Bonferroni correction across the restoration tests (Appendix A.4), establishing a behavioral contribution from tokens outside the dominant group.

Table 3: Early amplification blocks and the Gemma control on the shared 100-image SpatiaLab sample (98 eligible images for Qwen block 16). For each image, norm contrast is the selected tokens' mean norm divided by the unselected tokens' median norm. Values are medians across images at the intervention-block output. "Share" is the median clean selected-group fraction of the total squared distance from the activation mean. Qwen, LLaVA, and InternVL use the $10\times$-median rule; Phi and Gemma use the top 2.5%.

| Encoder | Block | Original | Attn. repl. | MLP repl. | Share |
|---|---|---|---|---|---|
| Qwen | 16 | 15.56 | 16.14 | 1.89 | .514 |
| Phi | 13 | 6.72 | 6.81 | 1.70 | .572 |
| LLaVA | 10 | 71.25 | 78.39 | 2.85 | .962 |
| Intern | 13 | 16.30 | 15.98 | 7.29 | .545 |
| Gemma | 14 | 1.71 | 1.71 | 1.69 | .080 |

# 5 MLP CONTRIBUTIONS TO HIGH-NORM TOKENS

The norm concentration in Section 4 motivates examining which encoder updates create and sustain the high-norm contrast. We use the component interventions defined in Section 3, with token-selection rules and controls detailed in Appendix E.

## 5.1 MLP UPDATES AMPLIFY THE NORM CONTRAST

We first measure the effect at the output of the intervened block. At the tested early amplification blocks in Qwen, Phi, LLaVA, and InternVL, MLP replacement substantially reduces the output norm contrast, whereas attention replacement has little effect or increases it (Table 3). The MLP update is therefore needed for the full measured contrast under this intervention. Attention replacement also changes the subsequent MLP input, so these tests do not isolate every indirect attention contribution or establish that the MLP alone is sufficient.

Applying the same replacement test at Gemma block 14 barely changes its norm contrast. This result is specific to the tested block; later-block controls, including Gemma block 27, are in Appendix Table A17.

## 5.2 EARLY MLP EFFECTS PERSIST THROUGH LATER BLOCKS

A local reduction does not establish whether later blocks restore the contrast. We therefore compare the downstream effects of early and later interventions near the encoder's end. In Phi, LLaVA, and InternVL, early MLP replacement still reduces the contrast, whereas replacing only the block-18 MLP changes it little (Appendix E.2, Table A18). Earlier MLP contributions therefore persist even where the tested later MLP contributes little locally.

Qwen differs. Replacing its block-16 MLP suppresses the contrast locally, but the contrast is larger than the clean baseline at block 31. Replacing the block-18 MLP instead produces a reduction that persists to block 31. The two interventions select different clean token groups. Thus, suppressing Qwen's first amplification does not establish lasting suppression.

## 5.3 DIFFERENT INTERNAL MLP PATHS CONTRIBUTE

The block-level tests identify MLP contributions but do not locate the amplification within each MLP. Figure 3 addresses this by tracing the selected and unselected groups through its internal stages. In Qwen, the norm contrast grows sharply at the SwiGLU product and remains large after the down projection. The displayed block 18 shows later amplification; Appendix Figure A7 also shows the earlier block-16 transition.

Mean-replacement tests confirm contributions from both Qwen branches. At block 18, gate replacement removes 12.6 percentage points more of the norm ratio than up-branch replacement (mean paired difference; unadjusted bootstrap 95% CI 11.8–13.4). Other encoders amplify the contrast through different paths: Phi's contrast develops across QuickGELU and FC2, while LLaVA and InternVL show amplification along GELU–FC2 paths. Their internal replacements reduce the early

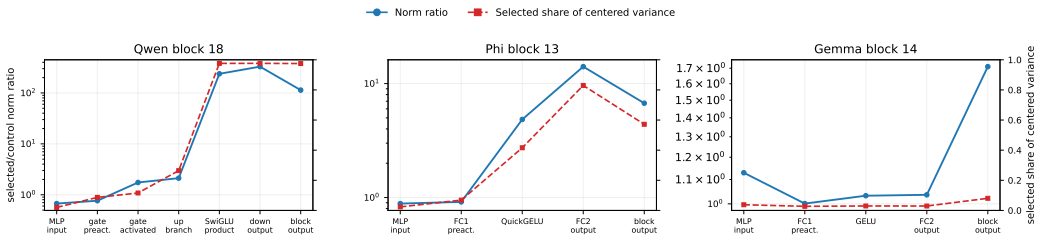



Figure 3: **Representative internal MLP paths on 100 SpatiaLab images.** Blue circles show the median selected/control norm ratio; red squares show the median selected-group share of centered variance. These are activation profiles, not intervention effects. MLP inputs are normalized, Qwen's gate and up branches are parallel, and only the final point includes residual addition. Left-axis ranges differ; every right axis spans zero to one.

contrast but have little effect at block 18 (Appendix E). Multiplicative gating is therefore not required for the observed amplification, and the tests do not establish that an activation function alone is sufficient.

These interventions test norm contrast; their effects on crosscoder fidelity are not measured.

## 6 Related Work

**Sparse features across layers and models.** Sparse autoencoders decompose dense activations into interpretable features (Bricken et al., 2023; Cunningham et al., 2023). Crosscoders extend this approach across layers and models (Lindsey et al., 2024). Their shared feature space supports tracking representations across depth and comparing model variants. However, sparsity penalties introduce artifacts in identifying features unique to a fine-tuned model (Minder et al., 2025). We examine how reconstruction scores characterize the token populations preserved by crosscoders within vision encoders.

**Reconstruction fidelity and model behavior.** Gao et al. (2024) evaluate sparse reconstructions through activation replacement, reporting favorable normalized errors for high-norm first-position language tokens despite substantial downstream disruption. SAEBench broadens evaluation to interpretability, disentanglement, and practical applications, showing that gains on proxy metrics do not reliably translate into practical improvements (Karvonen et al., 2025). In vision, Chatzoudis et al. (2026) evaluate cross-layer transcoders that reconstruct post-MLP activations in CLIP and measure their effect on classification. We extend this evaluation to token populations across VLM vision blocks and sparsity settings, making pooled EV's weighting explicit, separating it from population recentering, and testing behavioral consequences through block replacement.

**High-norm tokens and visual attention sinks.** Darcet et al. (2024) identify high-norm tokens in vision transformers and introduce registers to reduce feature-map artifacts. Jiang et al. (2025) identify sparse neurons that concentrate these activations and construct registers at inference time without retraining. In VLMs, Kang et al. (2025) improve performance by redistributing attention away from visual sinks. Luo et al. (2026) distinguish vision-encoder sinks from language-model-emerged sinks and show that propagated visual tokens carry high-level context with task-dependent benefits. These findings distinguish activation magnitude from behavioral importance. Preserving useful high-norm tokens is important, while fidelity elsewhere remains a separate evaluation question. We investigate how high-norm populations shape reconstruction fidelity and identify model-specific MLP contributions necessary for the measured norm contrasts, connecting activation structure to reconstruction errors and their behavioral effects.

## 7 Conclusion

In this work we showed that raw explained variance (EV) pools reconstruction across a token population it does not report, and that this pooling inflates apparent fidelity without establishing

typical-token fidelity or preserved VLM behavior. We found that raw and masked EV retain the same within-block dictionary ordering, meaning our results concern the interpretation of fidelity scores rather than which dictionary is best; and we showed that high activation norm alone does not establish that a token is behaviorally unimportant. This motivates evaluating crosscoder fidelity with masked, token-balanced, and norm-stratified scores alongside raw EV, and pairing important reconstructions with block replacement inside the original model, so that fidelity claims are tied to both the token population being scored and the behavior being explained.

**Limitations & future work.** Our evaluation used one seed per final setting, offline reconstructions inferred from the full clean activation stack, and crosscoders trained on CV-Bench Depth images, and was restricted to the tested models, blocks, and spatial tasks; it is therefore not certain that these findings generalize beyond this setting. Gemma lacks a downstream replacement control, our mask choices affect reported fidelity, and our t-SNE views are qualitative. Testing block replacement on general visual question answering and image classification would show whether the discrepancy between pooled EV and behavioral preservation extends beyond spatial reasoning, and training on broader image collections would help assess how the training distribution affects these findings. Multiple training seeds and a Gemma replacement control remain further tests we would be excited to see addressed.

## ETHICS STATEMENT

This study uses pretrained vision–language models and existing benchmark datasets, with no human-subject experiments or collection of private data. Our findings concern reconstruction fidelity and behavior on the evaluated datasets; they do not establish model safety, fairness, or reliability in deployment.

## REPRODUCIBILITY STATEMENT

Appendix B details activation locations, token handling, training settings, compute resources, intervention blocks, and statistical tests. Complete sweep tables appear in Appendix D.5, SpatiaLab results in Appendix C, and stress-test results in Appendix A. Upon acceptance, we will release the experiment code along with CV-Bench training split indices and SpatiaLab evaluation manifests; source images will be accessed through their original datasets and redistributed only where licenses permit.

## AI USE STATEMENT

Generative AI tools were mainly used for proofreading, enhancing text fluency, and checking grammar correctness. All experiment execution was carried out by the authors. The experimental design and framework are the original work of the authors.

## REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. URL https://arxiv.org/abs/2404.14219.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Gerasimos Chatzoudis, Konstantinos D. Polyzos, Zhuowei Li, Difei Gu, Gemma E. Moran, Hao Wang, and Dimitris N. Metaxas. Can cross-layer transcoders replace vision transformer activations? an interpretable perspective on vision, 2026. URL https://arxiv.org/abs/2604.13304.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024. URL https://arxiv.org/abs/2406.04093.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025. URL https://arxiv.org/abs/2506.08010.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models, 2025. URL https://arxiv.org/abs/2503.03321.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL https://arxiv.org/abs/2503.09532.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/crosscoders/.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models, 2026. URL https://arxiv.org/abs/2510.08510.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Robustly identifying concepts introduced during chat fine-tuning using crosscoders. *arXiv preprint arXiv:2504.02922*, 2025. URL https://arxiv.org/pdf/2504.02922v1.

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. *arXiv preprint arXiv:2406.16860*, 2024. URL https://arxiv.org/abs/2406.16860.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://www.jmlr.org/papers/v9/vandermaaten08a.html.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md. Iqramul Hoque, Shahriyar Zaman Ridoy, Mohammed Eunus Ali, Majd Hawasly, Mohammad Raza, and Md Rizwan Parvez. SpatiaLab: Can vision–language models perform spatial reasoning in the wild? In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fWWUPOb0CT.

# APPENDIX

The appendix begins with the supporting stress tests, followed by experimental details, complete SpatiaLab results, metric checks, mechanism controls, and separate norm-group visualizations for each encoder.

**Appendix contents**


| | | |
|---|---|---|
| A | Supporting threshold-free and norm-decile stress tests | 1 |
| B | Experimental details and reproducibility | 3 |
| C | Complete SpatiaLab evaluation | 5 |
| D | Metric decomposition and robustness | 7 |
| E | Mechanism interventions | 10 |
| F | Gemma control and final-block analysis | 16 |
| G | Norm-group visualizations across encoders | 17 |


## A SUPPORTING THRESHOLD-FREE AND NORM-DECILE STRESS TESTS

### A.1 DATASET AND SCORING

This targeted stress test uses 50 Visual Genome images held out from crosscoder training, validation, and checkpoint or sparsity selection. Training uses activations from 600 CV-Bench Depth images without answer supervision. The stress-test images were manually selected for two identifiable entities with a visually non-trivial front–behind relation; they do not represent general Visual Genome performance.

We annotated entity names, bounding boxes, and relations. Each image yields two entity-swapped questions using: "What is the spatial relationship between the [subject] and the [object] from the viewpoint of [object] in the image?" The final template resolves an earlier subject–reference ambiguity; all pairs were regenerated and checked for inverse labels. Preliminary scores are excluded. The final 100 queries contain 50 *Behind* and 50 *In Front* labels. Answers are scored from the two answer-letter logits, without a language-model judge. Inference uses 50 independent image pairs (Appendix B.4).

### A.2 COMPLETE BLOCK-REPLACEMENT RESULTS

Table A1 reports every dictionary and tested depth. Its EV column reproduces the late-block scores on the CV-Bench Depth validation split used for the controlled reconstruction sweep, whereas its behavioral columns use the separate 50-image stress test. Tables A2–A4 instead recompute fidelity directly on those 50 stress-test images. The stress-test images, prompts, answer choices, and statistical units also differ from SpatiaLab. Qwen's clean accuracy is .78 for *Behind* and .50 for *In Front*; pair-exact accuracy and clean-answer agreement therefore accompany aggregate accuracy.

### A.3 THRESHOLD-FREE FIDELITY

Tables A2–A4 report six-dictionary ranges recomputed on the 50-image stress-test activation sets. Qwen's block-24 highest-norm decile has EV .9991–.9997 across dictionaries, while typical-token EV spans $-.207$–.535. LLaVA also shows a compressed late raw-EV range. For InternVL, the wider raw-EV range provides a weaker example of compression. All metrics retain the within-block dictionary order.

Across the 18 repeated block–dictionary conditions, LLaVA's descriptive correlation with clean-answer agreement is $-.025$ for raw EV and .891 for typical-token EV. InternVL's token-balanced versus raw correlations are .934 versus .893 for agreement and .947 versus .861 for margin preservation. These are descriptive comparisons, not independent-sample significance tests. Qwen's equal-image-weighting check appears in Appendix D.7.

Table A1: Held-out binary-depth accuracy after block replacement. Early, middle, and late are 6/12/18 for Phi and LLaVA, 8/16/24 for Qwen, and 6/13/18 for InternVL. "CV-val EV" gives late-block raw and tower-specific union-masked EV on the CV-Bench Depth validation split (Appendix D.2); accuracy, clean agreement, and pair-exact accuracy are measured on the 50-image stress test. Clean accuracies are .68/.64/.68/.72 respectively.

| Tower | $\lambda$ | $h$ | CV-val EV raw/masked | Accuracy Early | Middle | Late | Late agree | Pair exact |
|---|---|---|---|---|---|---|---|---|
| Phi | .1 | 8192 | .974/.973 | .65 | .68 | .67 | .95 | .38 |
| Phi | .1 | 4096 | .970/.969 | .66 | .66 | .66 | .94 | .38 |
| Phi | .3 | 4096 | .950/.948 | .63 | .67 | .65 | .89 | .36 |
| Phi | 1 | 4096 | .899/.768 | .67 | .64 | .65 | .79 | .38 |
| Phi | 3 | 4096 | .845/.644 | .64 | .63 | .58 | .64 | .36 |
| Phi | 10 | 4096 | .775/.485 | .60 | .60 | .61 | .65 | .30 |
| Qwen | .1 | 4096 | .999/.755 | .66 | .65 | .66 | .90 | .36 |
| Qwen | .1 | 2048 | .998/.664 | .65 | .65 | .63 | .81 | .32 |
| Qwen | 1 | 4096 | .997/.467 | .68 | .62 | .54 | .70 | .10 |
| Qwen | 1 | 2048 | .997/.420 | .67 | .62 | .51 | .67 | .08 |
| Qwen | 10 | 4096 | .995/.065 | .61 | .47 | .52 | .68 | .06 |
| Qwen | 10 | 2048 | .995/.030 | .57 | .50 | .51 | .65 | .02 |
| LLaVA | .1 | 4096 | .995/.885 | .69 | .68 | .72 | .92 | .48 |
| LLaVA | .1 | 2048 | .994/.845 | .72 | .70 | .68 | .88 | .40 |
| LLaVA | 1 | 4096 | .990/.761 | .66 | .68 | .66 | .82 | .38 |
| LLaVA | 1 | 2048 | .988/.717 | .69 | .66 | .62 | .78 | .32 |
| LLaVA | 3 | 4096 | .987/.691 | .68 | .57 | .63 | .77 | .32 |
| LLaVA | 3 | 2048 | .985/.650 | .71 | .54 | .58 | .78 | .24 |
| Intern | .1 | 4096 | .955/.903 | .75 | .74 | .74 | .96 | .56 |
| Intern | .1 | 2048 | .939/.867 | .77 | .74 | .75 | .93 | .54 |
| Intern | 1 | 4096 | .887/.755 | .76 | .73 | .75 | .87 | .52 |
| Intern | 1 | 2048 | .871/.721 | .76 | .75 | .77 | .85 | .56 |
| Intern | 10 | 4096 | .769/.497 | .57 | .59 | .59 | .67 | .36 |
| Intern | 10 | 2048 | .756/.470 | .57 | .59 | .63 | .63 | .34 |

Table A2: Qwen EV ranges on the 50-image stress test across six dictionaries. "Img. typ." computes typical-token EV per image before averaging the images equally.

| Block | Raw | Token | Typical | Img. typ. | Acc. |
|---|---|---|---|---|---|
| 8 | .190–.859 | .188–.857 | .155–.850 | .148–.847 | .57–.68 |
| 16 | .732–.944 | .179–.823 | .116–.811 | .108–.808 | .47–.65 |
| 24 | .995–.998 | .834–.939 | −.207–.535 | −.218–.526 | .51–.66 |

Table A3: LLaVA EV ranges on the 50-image stress test across six dictionaries. "Dec." is norm-decile EV and "Typ." averages deciles 1–9.

| Block | Raw | Token | Dec. | Typ. | d10 |
|---|---|---|---|---|---|
| 6 | .448–.849 | .449–.849 | .441–.847 | .441–.847 | .447–.849 |
| 12 | .977–.990 | .446–.772 | .429–.765 | .366–.739 | .9967–.9986 |
| 18 | .973–.989 | .433–.761 | .382–.731 | .314–.701 | .9951–.9983 |

Table A4: InternVL EV ranges on the 50-image stress test across six dictionaries. "Dec." is norm-decile EV and "Typ." averages deciles 1–9.

| Block | Raw | Token | Dec. | Typ. | d10 |
|---|---|---|---|---|---|
| 6 | .401–.904 | .404–.905 | .390–.903 | .389–.903 | .407–.902 |
| 13 | .750–.944 | .257–.831 | .278–.836 | .204–.818 | .950–.989 |
| 18 | .647–.925 | .256–.843 | .268–.842 | .198–.827 | .900–.979 |

### A.4 TOKEN RESTORATION AND REPLACEMENT

Restoration inserts one clean norm decile into an otherwise reconstructed block output; replacement inserts one reconstructed decile into an otherwise clean output. Groups use clean activation norms. These interventions test contributions in different reconstruction contexts, not intrinsic importance.

**Qwen.** For the sparsest dictionary, full-reconstruction accuracy at blocks 8/16/24 is .57/.50/.51, versus .64 clean. Restoring decile 8 raises block-16 accuracy to .65; restoring decile 7 or 8 raises block-24 accuracy to .65. Decile-10 restoration reaches .58/.59 at blocks 16/24. The decile-8 gains over full reconstruction survive correction over ten tests (Bonferroni-adjusted image-clustered $p = .0027/.0131$). These tests establish a contribution outside the highest-norm group, not a significant decile-8 versus decile-10 difference.

**LLaVA and InternVL.** Table A5 gives the complete block-18 grids. Clean/full accuracies are .68/.58 for LLaVA and .72/.63 for InternVL. No isolated replacement reproduces the full loss. LLaVA's nominal restoration comparisons are not separate significance claims because ten groups were tested. InternVL's decile-9 gain over full reconstruction has unadjusted $p = .0192$ and Bonferroni-adjusted $p = .1921$; it also does not differ reliably from clean ($p = .8036$, 95% CI $[-.061, .101]$). InternVL's strict mask selects no tokens at block 6, 2427 (0.75%) at block 13, and 1227 (0.38%) at block 18.

Table A5: Block-18 accuracy after restoring or replacing one norm decile. Restoration starts from full reconstruction; replacement starts from clean activations.

| Model | Operation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA | Restore | .62 | .65 | .62 | .65 | .65 | .66 | .67 | .65 | .66 | .65 |
| LLaVA | Replace | .68 | .69 | .68 | .68 | .70 | .70 | .68 | .70 | .68 | .68 |
| InternVL | Restore | .65 | .70 | .71 | .70 | .70 | .71 | .73 | .73 | .74 | .71 |
| InternVL | Replace | .72 | .72 | .71 | .71 | .72 | .72 | .72 | .69 | .72 | .73 |

### A.5 IMAGE-CLUSTERED COMPARISONS

Table A6 applies Holm correction across the six highlighted sparse-onset comparisons. Restoration grids instead use Bonferroni correction across ten tests; methods are in Appendix B.4. This audit uses saved outputs without new inference.

Table A6: Image-clustered sparse-onset comparisons. $\Delta$ is replacement accuracy minus unreplaced accuracy; $p_H$ is Holm-adjusted over all six rows.

| Encoder | Condition | $\Delta$ | 95% CI | $p$ | $p_H$ |
|---|---|---|---|---|---|
| Qwen | b16, 4096 | $-.17$ | $[-.254, -.086]$ | .0005 | .0029 |
| Qwen | b16, 2048 | $-.14$ | $[-.231, -.049]$ | .0066 | .0331 |
| LLaVA | b12, 4096 | $-.11$ | $[-.222, .002]$ | .0800 | .0800 |
| LLaVA | b12, 2048 | $-.14$ | $[-.244, -.036]$ | .0153 | .0460 |
| InternVL | b13, 4096 | $-.13$ | $[-.237, -.023]$ | .0284 | .0567 |
| InternVL | b13, 2048 | $-.13$ | $[-.220, -.040]$ | .0100 | .0399 |

## B EXPERIMENTAL DETAILS AND REPRODUCIBILITY

### B.1 ENCODERS, ACTIVATION EXTRACTION, AND TRAINING

We collect residual activations at each vision block from 600 images drawn from the CV-Bench Depth subset, using the same sampled token positions across blocks for each image. Crosscoder training uses these activations without task labels or answer supervision. For Phi, we retain only the first 23 blocks of its 24-block CLIP encoder because the VLM consumes the penultimate block's

Table A7: Vision encoders used in the study. Width is the vision residual width. "Live" excludes a block whose output is not consumed downstream.

| Encoder | Live blocks | Width |
|---|---|---|
| Qwen2.5-VL-3B-Instruct | 32 | 1280 |
| LLaVA-OneVision-1.5-4B | 24 | 1024 |
| InternVL2.5-4B | 24 | 1024 |
| Phi-3.5-Vision | 23 | 1024 |
| Gemma-3-4B | 27 | 1152 |

Table A8: Shared settings for final crosscoder sweeps.

| Setting | Value |
|---|---|
| optimizer | Adam, $\beta_1 = .9$, $\beta_2 = .999$ |
| learning rate | $5 \times 10^{-5}$ |
| batch size | 1024 tokens |
| epochs | 20 |
| validation split | 15%, split by image |
| maximum sampled tokens/image | 1024 for dynamic-token towers |
| encoder/decoder computation | fp32 |
| saved activations | bfloat16 |
| learning-rate schedule | constant to 80%, then linear decay |
| $\ell_1$ warm-up | first 5% of steps |
| gradient clipping | max norm 1.0 |

output. The unused final block is excluded from both crosscoder training and intervention. Table A8 summarizes the shared training settings and numerical precision.

Block $b$ is scaled by $\sqrt{d}/\bar{n}_b$, where $d$ is residual width and $\bar{n}_b$ is that block's mean training-token norm. Evaluation reuses the training scales and image split; recomputed raw EV must match the training record within $5 \times 10^{-3}$. Final sweeps use one initialization seed per setting. Short seed probes assess pipeline stability, not variation across fully trained models.

**Crosscoder objective.** For a minibatch of $N$ aligned tokens, let $\tilde{x}_{ib}$ be token $i$'s scaled activation at block $b$. The shared ReLU code and block reconstruction are $z_i = \mathrm{ReLU}(\sum_b E_b \tilde{x}_{ib} + c)$ and $\hat{\tilde{x}}_{ib} = D_b z_i + d_b$, where $E_b, D_b$ are learned encoder and decoder matrices and $c, d_b$ are biases. Training minimizes

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left[ \sum_b \|\tilde{x}_{ib} - \hat{\tilde{x}}_{ib}\|_2^2 + \lambda \sum_k z_{ik} \sum_b \|D_{b,:,k}\|_2 \right]. \tag{5}$$

Here, $k$ indexes dictionary features and $D_{b,:,k}$ is feature $k$'s decoder vector at block $b$. The loss averages over tokens and sums over blocks and activation dimensions. Decoder vectors are initialized with norm .08 and then learned without a fixed unit-norm constraint; their norms enter the sparsity penalty. Table A8 specifies the $\ell_1$ warm-up.

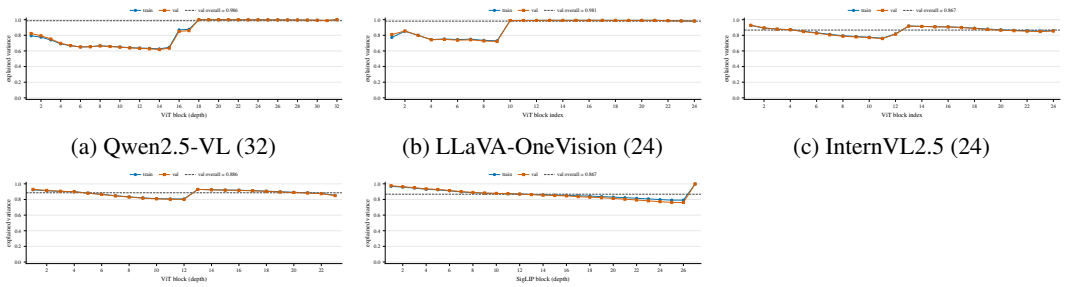
### B.2 Training diagnostics and compute

All runs use one NVIDIA RTX A5000 Laptop GPU (16 GB VRAM) and 128 GB host RAM. Activations are processed in chunks, and one VLM is loaded at a time during scoring. Compute and storage constraints limit replication across training seeds.

### B.3 SpatiaLab protocol

We evaluate all 1,400 SpatiaLab MCQs across six categories and 30 subcategories (Wasi et al., 2026). Category counts appear in Table A9.

The Checkpoints, blocks, masks, prompts, and statistics were frozen before inspecting replacement results. Preprocessing matches activation collection: Qwen's slow processor, four Phi crops, LLaVA's



(a) Qwen2.5-VL (32) (b) LLaVA-OneVision (24) (c) InternVL2.5 (24)

(d) Phi-3.5-Vision (23) (e) Gemma-3 (27)

Figure A1: Per-block training and validation EV for one multi-block crosscoder per tower. Tower depths differ, so block indices are not aligned.

fast processor, and at most six dynamic InternVL tiles. A clean pass records answer scores and block statistics. Reconstructions use the full clean activation stack and replace one output at a time: Qwen blocks 16/24, Phi and LLaVA 12/18, or InternVL 13/18. We compare the resulting answer scores with the clean baseline.

**Stress-test execution.** Before scoring cached reconstructions, we verify their dimensions, numerical precision, checkpoint, block, and image identity. Replacing an activation with its clean value must leave the scored answer logits exactly unchanged.

### B.4 Statistical inference

**SpatiaLab.** For each model's 12 planned dictionary–block comparisons, we use an exact two-sided sign-flip test, 20,000-draw paired image-bootstrap 95% intervals, and Holm correction within the model. Each image has one question. Per-category results are descriptive and do not guide checkpoint selection.

**Stress test.** Each of the 50 images contributes two entity-swapped questions. We compute intervention-minus-reference correctness for each question (correct $= 1$, incorrect $= 0$), then average the two differences per image. The reference is the clean or full-reconstruction condition specified for each comparison. An exact two-sided clustered sign-flip test reverses both question differences together within each image, treating images as the independent units. Student-$t$ 95% confidence intervals for the mean paired difference use the 50 image averages and are unadjusted. Holm correction covers the six comparisons in Table A6; Bonferroni correction covers each ten-decile restoration grid.

**Availability.** Upon acceptance, we will release code, CV-Bench split indices, SpatiaLab manifests, Visual Genome IDs, stress-test annotations, prompts, and preprocessing manifests. Source images remain subject to their original licenses and are redistributed only where permitted.

## C Complete SpatiaLab evaluation

### C.1 Configuration and task coverage

Each VLM uses six dictionaries at two blocks under Appendix B.3's protocol. Phi's additional dense behavioral control ($h = 8192$, $\lambda = .1$) is outside its fixed-width controlled sweep; the four controlled sweeps therefore contain 23 cells. Table A9 expands the main overall-accuracy table. Task comparisons are descriptive; significance markers apply only to overall accuracy.

### C.2 Answer-distribution changes

We compute $D_{\mathrm{KL}}(p_{\mathrm{clean}} \| p_{\mathrm{reconstructed}})$ from each image's four answer-letter logits. At width 4096 and the late block, dense/intermediate/sparse settings use $\lambda = .1, 1, 10$ for Qwen, Phi, and InternVL, and $.1, 1, 3$ for LLaVA. Both the center and right tail shift with sparsity (Figure A2).

Table A9: SpatiaLab accuracy (%) by task. Each panel reports the clean model and six frozen dictionaries at two blocks. All four models use all 1,400 images; the category counts are 238/259/202/212/252/237 in the order shown. $B$ is the replacement block and $d$ is dictionary width. † marks an overall accuracy loss relative to the clean model that remains nonzero after Holm correction across the 12 comparisons within that model ($p < .05$). The Qwen panel uses the base checkpoint rerun.

**Qwen2.5-VL-3B-Instruct**

| $B$ | $\lambda$ | $d$ | 3D Geometry | Depth & Occlusion | Orientation | Relative Positioning | Size & Scale | Spatial Navigation | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Clean | – | – | 42.9 | 51.7 | 47.5 | 59.0 | 45.6 | 44.7 | 48.4 |
| 16 | .1 | 2048 | 47.9 | 49.4 | 48.5 | 56.6 | 44.8 | 41.8 | 48.0 |
| 16 | .1 | 4096 | 44.5 | 51.0 | 47.5 | 56.1 | 43.7 | 45.6 | 47.9 |
| 16 | 1 | 2048 | 43.3 | 46.7 | 48.0 | 54.2 | 42.1 | 41.4 | 45.7 |
| 16 | 1 | 4096 | 41.6 | 47.9 | 48.0 | 53.8 | 40.9 | 43.0 | 45.6 |
| 16 | 10 | 2048 | 42.4 | 40.9 | 50.5 | 47.2 | 39.7 | 40.9 | 43.3† |
| 16 | 10 | 4096 | 42.0 | 40.2 | 48.0 | 49.1 | 41.7 | 40.5 | 43.3† |
| 24 | .1 | 2048 | 45.8 | 51.0 | 50.0 | 57.1 | 42.9 | 43.5 | 48.1 |
| 24 | .1 | 4096 | 43.7 | 51.7 | 49.5 | 61.3 | 42.1 | 43.0 | 48.3 |
| 24 | 1 | 2048 | 42.9 | 47.1 | 48.5 | 42.0 | 38.1 | 40.9 | 43.1† |
| 24 | 1 | 4096 | 41.6 | 46.7 | 49.5 | 43.9 | 37.7 | 39.7 | 43.0† |
| 24 | 10 | 2048 | 45.0 | 38.6 | 48.5 | 41.5 | 39.7 | 37.1 | 41.5† |
| 24 | 10 | 4096 | 43.7 | 42.1 | 49.0 | 41.0 | 41.3 | 40.5 | 42.8† |

**Phi-3.5-Vision-Instruct**

| $B$ | $\lambda$ | $d$ | 3D Geometry | Depth & Occlusion | Orientation | Relative Positioning | Size & Scale | Spatial Navigation | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Clean | – | – | 46.2 | 43.2 | 44.1 | 47.6 | 45.6 | 50.2 | 46.1 |
| 12 | .1 | 4096 | 45.4 | 42.1 | 44.1 | 49.5 | 45.2 | 49.8 | 45.9 |
| 12 | .1 | 8192 | 46.2 | 43.2 | 43.6 | 50.0 | 44.4 | 49.8 | 46.1 |
| 12 | .3 | 4096 | 45.8 | 42.5 | 44.1 | 47.2 | 45.2 | 49.8 | 45.7 |
| 12 | 1 | 4096 | 46.2 | 40.9 | 41.1 | 47.2 | 45.2 | 45.6 | 44.4 |
| 12 | 3 | 4096 | 44.5 | 39.0 | 38.1 | 42.5 | 46.4 | 45.1 | 42.7† |
| 12 | 10 | 4096 | 44.1 | 39.0 | 40.1 | 36.8 | 42.9 | 37.6 | 40.1† |
| 18 | .1 | 4096 | 45.8 | 44.0 | 43.6 | 49.5 | 44.4 | 50.2 | 46.2 |
| 18 | .1 | 8192 | 45.4 | 43.6 | 44.1 | 50.0 | 45.6 | 49.8 | 46.4 |
| 18 | .3 | 4096 | 47.5 | 42.1 | 44.6 | 48.1 | 44.4 | 51.1 | 46.2 |
| 18 | 1 | 4096 | 42.0 | 45.2 | 41.1 | 45.8 | 44.4 | 48.1 | 44.5 |
| 18 | 3 | 4096 | 43.3 | 40.5 | 37.6 | 34.0 | 43.7 | 41.8 | 40.4† |
| 18 | 10 | 4096 | 45.8 | 36.7 | 32.7 | 33.0 | 41.3 | 38.4 | 38.2† |

**LLaVA-OneVision-1.5-4B-Instruct**

| $B$ | $\lambda$ | $d$ | 3D Geometry | Depth & Occlusion | Orientation | Relative Positioning | Size & Scale | Spatial Navigation | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Clean | – | – | 50.8 | 52.9 | 54.5 | 55.7 | 53.6 | 50.2 | 52.9 |
| 12 | .1 | 2048 | 49.6 | 53.7 | 54.5 | 54.2 | 53.6 | 51.5 | 52.8 |
| 12 | .1 | 4096 | 48.3 | 54.1 | 54.0 | 57.1 | 53.2 | 51.5 | 52.9 |
| 12 | 1 | 2048 | 46.2 | 47.9 | 49.5 | 51.4 | 52.4 | 51.1 | 49.7† |
| 12 | 1 | 4096 | 49.2 | 49.4 | 49.5 | 55.2 | 53.2 | 52.7 | 51.5 |
| 12 | 3 | 2048 | 46.2 | 43.2 | 46.5 | 44.3 | 53.2 | 48.5 | 47.1† |
| 12 | 3 | 4096 | 47.1 | 39.8 | 44.6 | 47.6 | 51.6 | 48.5 | 46.5† |
| 18 | .1 | 2048 | 47.9 | 51.7 | 51.0 | 56.1 | 53.2 | 51.9 | 51.9 |
| 18 | .1 | 4096 | 49.2 | 54.4 | 53.5 | 55.7 | 53.6 | 51.5 | 52.9 |
| 18 | 1 | 2048 | 47.5 | 46.7 | 49.0 | 48.6 | 54.4 | 50.6 | 49.5† |
| 18 | 1 | 4096 | 44.5 | 47.5 | 50.5 | 56.6 | 54.8 | 51.1 | 50.7 |
| 18 | 3 | 2048 | 45.4 | 44.0 | 48.5 | 44.8 | 52.4 | 51.5 | 47.8† |
| 18 | 3 | 4096 | 47.5 | 45.9 | 49.0 | 45.8 | 53.2 | 54.0 | 49.3† |

**InternVL2.5-4B**

| $B$ | $\lambda$ | $d$ | 3D Geometry | Depth & Occlusion | Orientation | Relative Positioning | Size & Scale | Spatial Navigation | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Clean | – | – | 49.2 | 45.2 | 51.5 | 57.5 | 45.6 | 52.7 | 50.0 |
| 13 | .1 | 2048 | 50.0 | 47.9 | 50.0 | 56.1 | 45.6 | 53.2 | 50.3 |
| 13 | .1 | 4096 | 49.2 | 45.2 | 51.5 | 58.0 | 46.4 | 54.0 | 50.4 |
| 13 | 1 | 2048 | 49.6 | 42.5 | 50.5 | 53.8 | 44.8 | 49.8 | 48.2 |
| 13 | 1 | 4096 | 50.0 | 45.6 | 50.0 | 53.3 | 45.2 | 51.1 | 49.0 |
| 13 | 10 | 2048 | 48.3 | 40.2 | 45.5 | 44.3 | 44.0 | 41.8 | 43.9† |
| 13 | 10 | 4096 | 49.6 | 38.2 | 46.0 | 42.9 | 42.5 | 43.0 | 43.6† |
| 18 | .1 | 2048 | 50.8 | 45.6 | 50.0 | 56.1 | 44.0 | 52.3 | 49.6 |
| 18 | .1 | 4096 | 50.0 | 45.2 | 51.0 | 55.7 | 44.0 | 52.7 | 49.5 |
| 18 | 1 | 2048 | 47.9 | 45.2 | 47.5 | 56.1 | 43.3 | 51.9 | 48.4 |
| 18 | 1 | 4096 | 52.5 | 44.8 | 49.0 | 57.1 | 44.8 | 51.5 | 49.7 |
| 18 | 10 | 2048 | 50.0 | 37.8 | 47.5 | 45.8 | 43.7 | 45.1 | 44.8† |
| 18 | 10 | 4096 | 48.7 | 41.3 | 46.0 | 50.0 | 43.7 | 47.3 | 46.0† |

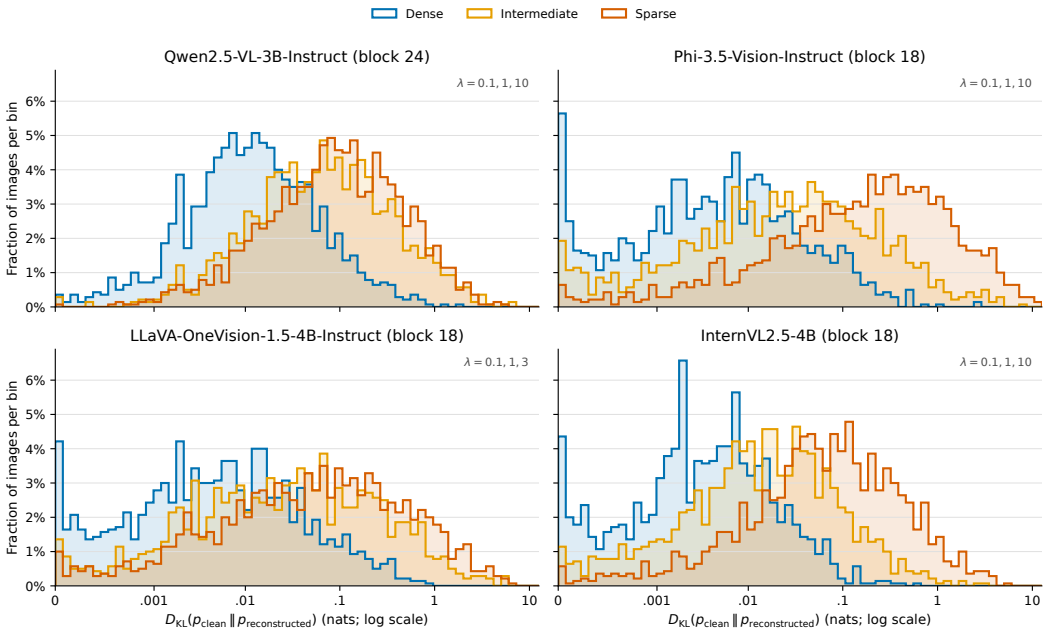



Figure A2: Per-image answer-distribution change after block replacement on all 1,400 SpatiaLab images. Each panel compares width-4096 dense, intermediate, and sparse crosscoders at the late evaluated block. The horizontal axis shows $D_{\mathrm{KL}}(p_{\mathrm{clean}} \,\|\, p_{\mathrm{reconstructed}})$ in nats on a logarithmic scale; the vertical axis is the fraction of images in each shared bin.

# D METRIC DECOMPOSITION AND ROBUSTNESS

## D.1 NUMERICAL FLOOR FOR TOKEN-BALANCED EV

For each block and evaluation population, we set $\epsilon = \max(10^{-8}\,\mathrm{median}_{i\in S}\, v_i, \tau)$, using the lower middle denominator when the token count is even. Here, $\tau$ is the smallest positive normal value in FP64 for the SpatiaLab and BatchTopK evaluators, and in FP32 for the separate threshold-free stress-test and Gemma evaluators. The saved epsilon values in these audits exceed both fallback floors, so this precision distinction does not change the selected epsilon. Means and denominators are computed from clean activations.

## D.2 MASK DEFINITIONS AND SENSITIVITY

Scores follow Section 2. Table A10 uses a cross-depth union of tokens above $10\times$ the block median. Controlled Qwen, LLaVA, and InternVL sweeps instead use the union of each block's top 2.5%. Phi's earlier sweep uses norm $> 100$, following the ViT criterion in Luo et al. (2026) and Darcet et al. (2024); this absolute threshold is specific to Phi's activation scale.

Table A10: Five-encoder audit under the common $10\times$-median union rule.

| Vision encoder | Blocks | Raw | Masked | Gap |
|---|---|---|---|---|
| Qwen2.5-VL | 32 | .986 | .663 | .323 |
| LLaVA-OneVision | 24 | .981 | .761 | .220 |
| InternVL2.5 | 24 | .867 | .800 | .067 |
| Phi-3.5-Vision | 23 | .880 | .877 | .003 |
| Gemma-3 | 27 | .867 | .867 | .000 |

The full-depth SpatiaLab audit (Figure 2) streams means, medians, and 90th-percentile cutoffs in one pass, then errors and retained-population denominators in a second. The $10\times$ mask is a fixed cross-depth union; D10 masking is block-local and retains approximately 90% of tokens. Positive-model plots use width 4096 and $\lambda = 1$.

Table A11: Phi threshold robustness at $\lambda = 10$, width 4096.

| Criterion | EV | Drop from raw |
|---|---|---|
| none (raw) | .705 | — |
| $10\times$ median | .697 | .008 |
| top .5% | .650 | .055 |
| top 1% | .616 | .089 |
| $5\times$ median | .553 | .152 |
| top 2.5% | .543 | .161 |
| norm $> 100$ | .527 | .178 |
| $3\times$ median | .513 | .191 |
| top 5% | .495 | .210 |

Phi's reconstruction gap depends on the selected population: the strict $10\times$-median rule removes few tokens and barely changes aggregate EV, whereas broader masks reveal a larger discrepancy (Table A11). Threshold-free checks are reported in Appendix A.3.

## D.3 ERROR AND VARIANCE ACROSS DEPTH TRANSITIONS

For each block, let $x_i$ and $\hat{x}_i$ be the clean and reconstructed activations of its $N$ evaluated tokens, with clean mean $\mu = N^{-1}\sum_i x_i$. SSE is the summed squared reconstruction error; TSS is the total squared deviation of clean activations from their mean:

$$\mathrm{SSE} = \sum_i \|x_i - \hat{x}_i\|_2^2, \qquad \mathrm{TSS} = \sum_i \|x_i - \mu\|_2^2.$$

Thus, $\mathrm{EV}_{\mathrm{raw}} = 1 - \mathrm{SSE}/\mathrm{TSS}$ (Equation 1). Table A12 compares $\mathrm{SSE}/N$ and $\mathrm{TSS}/N$ across selected depth transitions. Values use saved full-depth reference statistics and EV outputs, with SSE recovered as $(1 - \mathrm{EV}_{\mathrm{raw}})\mathrm{TSS}$.

Table A12: Error and variance changes on all 1,400 SpatiaLab images, using the same crosscoders as Figure 2. Ratios divide the later block's per-token value by the earlier block's value. These are SpatiaLab results, separate from the 50-image audits.

| Transition | TSS/token ratio | SSE/token ratio | Raw EV |
|---|---|---|---|
| Qwen $31 \to 32$ | $32.1\times$ | $3.4\times$ | $.982 \to .998$ |
| Gemma $26 \to 27$ | $940.9\times$ | $5.4\times$ | $.444 \to .997$ |
| Phi $12 \to 13$ | $2.29\times$ | $1.05\times$ | $.668 \to .847$ |
| Phi $13 \to 23$ | $1.43\times$ | $2.81\times$ | $.847 \to .700$ |

Qwen's and Gemma's final-block EV increases occur because centered variance grows faster than reconstruction error, despite an increase in absolute error. Phi shows the same relationship at block $12 \to 13$, followed by faster error growth than variance growth over blocks $13 \to 23$. These comparisons explain the score changes, without identifying their architectural causes or establishing changes in encoder capability.

### D.4 Population recentering

Table A13 fixes tokens and reconstruction errors to the lower nine block-local norm deciles, changing only the denominator's center (Equation 4). It separates recentering from token selection and unequal weighting.

Table A13: Centering comparison for the fixed lower-nine-decile population at representative late blocks. $Q^{(g)}$ uses the block-wide mean, $Q^{(\ell)}$ uses the retained group's mean, and "Between" is $B_D/(W_D + B_D)$.

| Encoder | Block | $Q^{(g)}$ | $Q^{(\ell)}$ | Between |
|---|---|---|---|---|
| Qwen2.5-VL | 24 | .890 | .225 | .858 |
| Phi-3.5-Vision | 18 | .571 | .559 | .027 |
| LLaVA-OneVision | 18 | .570 | .531 | .083 |
| InternVL2.5 | 18 | .614 | .613 | .004 |
| Gemma-3 | 26 | .456 | .455 | .001 |

### D.5 Complete controlled sweeps

Raw, masked, and typical-token EV preserve the same complete dictionary ordering within each tower. The result is a growing difference in score scale, not a reversal of the preferred dictionary.

### D.6 Single-block BatchTopK control

We train independent BatchTopK sparse autoencoders (Bussmann et al., 2024) on Qwen block 24, selected because Qwen exhibits the strongest crosscoder discrepancy. All three dictionaries have width 4096 and use the same 600-image CV-Bench Depth corpus, 510/90 image-level training/validation split (seed 49), and up to 1024 sampled tokens per image as the Qwen crosscoder. A single scale factor, $\sqrt{1280}$ divided by the mean training-token norm, preserves relative token magnitudes; no per-token normalization is applied.

We train each SAE for 20 epochs with dictionary width 4096, batch size 1024, and initialization seed 0. Adam uses learning rate $3 \times 10^{-4}$ and $\beta = (.9, .99)$. Training uses no $\ell_1$ penalty, unit-norm decoder rows, and the BatchTopK auxiliary loss. Evaluation uses fixed validation batches because feature selection operates across tokens within each batch. Checkpoints are selected by validation raw EV; all three selected checkpoints occur at epoch 20.

Table A14: All final controlled sweep cells. "Typical" is the mean EV over norm deciles 1–9. Masks are tower-specific (Appendix D.2).

| Tower | $\lambda$ | $h$ | $\ell_0$ | Raw | Masked | Typical |
|---|---|---|---|---|---|---|
| Phi | .1 | 4096 | 2241.0 | .959 | .957 | .922 |
| Phi | .3 | 4096 | 1093.2 | .934 | .932 | .876 |
| Phi | 1 | 4096 | 220.7 | .880 | .808 | .778 |
| Phi | 3 | 4096 | 57.5 | .812 | .699 | .657 |
| Phi | 10 | 4096 | 26.3 | .705 | .527 | .459 |
| Qwen | .1 | 2048 | 1357.3 | .994 | .848 | .821 |
| Qwen | .1 | 4096 | 2339.2 | .996 | .887 | .865 |
| Qwen | 1 | 2048 | 127.6 | .984 | .594 | .526 |
| Qwen | 1 | 4096 | 196.8 | .986 | .649 | .588 |
| Qwen | 10 | 2048 | 44.9 | .973 | .315 | .197 |
| Qwen | 10 | 4096 | 57.9 | .975 | .352 | .236 |
| LLaVA | .1 | 2048 | 1441.6 | .989 | .858 | .845 |
| LLaVA | .1 | 4096 | 2404.2 | .991 | .889 | .878 |
| LLaVA | 1 | 2048 | 174.1 | .977 | .706 | .686 |
| LLaVA | 1 | 4096 | 242.2 | .981 | .752 | .733 |
| LLaVA | 3 | 2048 | 59.9 | .969 | .603 | .578 |
| LLaVA | 3 | 4096 | 76.2 | .973 | .645 | .620 |
| Intern | .1 | 2048 | 1483.7 | .922 | .875 | .865 |
| Intern | .1 | 4096 | 2611.3 | .945 | .912 | .904 |
| Intern | 1 | 2048 | 226.5 | .848 | .755 | .741 |
| Intern | 1 | 4096 | 282.9 | .867 | .784 | .769 |
| Intern | 10 | 2048 | 21.3 | .660 | .451 | .422 |
| Intern | 10 | 4096 | 23.3 | .677 | .476 | .446 |

Table A15: Single-block BatchTopK fidelity on Qwen block 24. All rows use the same 92,160 sampled tokens from 90 CV-Bench Depth validation images. Realized mean $\ell_0$ equals the target batch budget $k$. Scores use original activation units and the definitions in Section 2: "Token" is token-balanced EV, "Typical" averages the separately recentered EVs of norm deciles 1–9, and "D10-masked" removes decile 10 and recenters the remaining tokens together.

| Mean $\ell_0$ | Raw | Token | Typical | D10-masked |
|---|---|---|---|---|
| 1024 | .99992 | .99727 | .97615 | .98614 |
| 256 | .99921 | .97029 | .79606 | .83256 |
| 64 | .99862 | .94856 | .66843 | .70622 |

As sparsity increases, typical-token EV falls substantially while raw EV remains above .998 (Table A15); all metrics retain the same dictionary ordering. The highest-norm decile holds 97.67% of globally centered variance. As in the crosscoder analysis, differences involving masked and typical-token scores reflect population selection and recentering as well as weighting. This single-seed, single-block control establishes the discrepancy without an explicit $\ell_1$ penalty or a shared multi-block code. Its optimizer settings and architecture differ from the crosscoder; it does not establish a matched performance comparison or behavioral effects.

### D.7 Equal image weighting

Qwen yields 216–3456 visual tokens per stress-test image. Each image-balanced metric uses the same global block mean and decile boundaries, then averages the 50 image scores equally (Table A16). Raw EV remains saturated, typical-token EV spans a broad range, and the within-block ordering is unchanged. Across 18 repeated conditions, image-balanced typical-token EV has descriptive correlations .873/.960/.986 with accuracy/agreement/margin preservation; image-balanced raw EV gives $-.039/-.092/-.051$. These are not independent-sample tests.

Table A16: Qwen block-24 token-weighted and image-balanced EV. "TB" is token-balanced EV and "Typ." averages norm deciles 1–9.

| $\lambda, h$ | Raw | Img. raw | TB | Img. TB | Typ. | Img. typ. |
|---|---|---|---|---|---|---|
| .1,4096 | .9981 | .9980 | .9387 | .9360 | .5353 | .5256 |
| .1,2048 | .9975 | .9974 | .9197 | .9164 | .4051 | .3928 |
| 1,4096 | .9965 | .9964 | .8880 | .8842 | .1929 | .1788 |
| 1,2048 | .9963 | .9962 | .8830 | .8793 | .1634 | .1497 |
| 10,4096 | .9950 | .9949 | .8362 | .8341 | −.1885 | −.1972 |
| 10,2048 | .9949 | .9948 | .8338 | .8316 | −.2072 | −.2180 |

## E Mechanism interventions

### E.1 Protocol and source replacement

All five encoders use the same 100 SpatiaLab images sampled uniformly without replacement from the 1,400-image benchmark. Selection uses Python's `random.Random(2027).sample` after sorting item IDs as strings. The frozen manifest records ordered image IDs, image and annotation hashes, and prompts; images are not selected using activation or intervention results. This mechanism sample is separate from the 50-image Visual Genome behavioral stress test in Appendix A.

Qwen, LLaVA, and InternVL select tokens above $10\times$ the clean intervention-block within-image median norm; Phi and Gemma select the top 2.5%. Each intervention replaces selected-token writes with the same image's unselected-token mean. The selected token indices remain fixed across conditions and downstream readout blocks. Per-image norm contrast is the selected tokens' mean norm divided by the unselected tokens' median norm; "Share" is their fraction of squared distance from the all-token activation mean. Tables report medians across eligible images unless stated otherwise. The residual identity $x_{\text{out}} = x_{\text{in}} + a + m$ and reconstructed internal MLP functions have zero maximum reported relative error. Attention replacement also changes the subsequent MLP input, so interventions test necessity and are paired with observational write decompositions.

Two Qwen images at block 16 have no qualifying tokens and are excluded from its local and propagated intervention statistics without replacement ($n = 98$). All other tested block–model conditions have $n = 100$. Table A17 reports early and later local effects. Local attention-versus-MLP and Qwen branch comparisons use 10,000 paired image-bootstrap draws for the mean difference (seed 2027 for local source comparisons; 2028 for branch comparisons). Intervals are unadjusted 95% percentile intervals, not simultaneous confidence statements across conditions.

Table A17: Attention-versus-MLP mean replacement on the 100-image SpatiaLab mechanism sample. Values are median selected/control norm ratios at the intervention-block output. "Share" is the median clean selected-group fraction of centered variance; $n$ counts eligible images.

| Encoder | Block | $n$ | Clean | Attn. repl. | MLP repl. | Share |
|---|---|---|---|---|---|---|
| Qwen | 16 | 98 | 15.56 | 16.14 | 1.89 | 0.514 |
| Qwen | 18 | 100 | 113.80 | 126.70 | 5.41 | 0.975 |
| Phi | 13 | 100 | 6.72 | 6.81 | 1.70 | 0.572 |
| Phi | 18 | 100 | 5.54 | 5.54 | 5.50 | 0.483 |
| LLaVA | 10 | 100 | 71.25 | 78.39 | 2.85 | 0.962 |
| LLaVA | 18 | 100 | 70.23 | 70.24 | 70.09 | 0.955 |
| InternVL | 13 | 100 | 16.30 | 15.98 | 7.29 | 0.545 |
| InternVL | 18 | 100 | 12.76 | 12.73 | 12.71 | 0.306 |
| Gemma | 14 | 100 | 1.71 | 1.71 | 1.69 | 0.080 |
| Gemma | 27 | 100 | 1.40 | 1.16 | 0.82 | 0.035 |

### E.2 Propagated effects near the encoder's end

Table A18 measures the same selected tokens after all intervening blocks run normally. Readouts use the penultimate consumed block (Qwen 31, Phi 22, LLaVA/InternVL 23), separating persistence from terminal block effects. For each image, $\Delta_j = R_j^{\text{clean}} - R_j^{\text{MLP}}$ is the reduction in selected/control norm ratio at the readout block. We report its mean and a paired image-bootstrap percentile 95% interval using 10,000 resamples independently for each condition. These intervals are unadjusted; they describe the mean paired effect, not uncertainty around the condition medians in the preceding columns.

Table A18: Propagated source interventions on the shared SpatiaLab sample. $b \to r$ denotes intervention and downstream readout blocks. Clean, attention, and MLP columns report median norm ratios. The last column reports the *mean paired* clean-minus-MLP reduction and its unadjusted 95% bootstrap interval. Token groups are fixed within an intervention but can differ between intervention blocks, so clean baselines need not coincide.

| Encoder | $b \to r$ | $n$ | Clean | Attn. | MLP | $\bar{\Delta}$ [95% CI] |
|---|---|---|---|---|---|---|
| Qwen | 16→31 | 98 | 31.74 | 31.41 | 37.02 | $-4.604$ $[-6.082, -3.042]$ |
| Qwen | 18→31 | 100 | 23.92 | 26.63 | 1.49 | $22.010$ $[21.508, 22.494]$ |
| Phi | 13→22 | 100 | 4.92 | 4.97 | 1.46 | $3.253$ $[3.068, 3.428]$ |
| Phi | 18→22 | 100 | 4.93 | 4.93 | 4.90 | $0.045$ $[0.040, 0.051]$ |
| LLaVA | 10→23 | 100 | 15.73 | 17.03 | 1.97 | $13.780$ $[13.458, 14.083]$ |
| LLaVA | 18→23 | 100 | 15.53 | 15.55 | 15.49 | $0.034$ $[0.031, 0.037]$ |
| InternVL | 13→23 | 100 | 6.49 | 6.40 | 3.21 | $3.366$ $[3.306, 3.430]$ |
| InternVL | 18→23 | 100 | 7.25 | 7.23 | 7.22 | $0.028$ $[0.024, 0.031]$ |

In Phi, LLaVA, and InternVL, early MLP replacement reduces contrast near the encoder's end, whereas block-18 replacement changes it little. This supports persistence of earlier MLP contributions under the tested interventions. Qwen differs: block-18 replacement reduces the block-31 median ratio from 23.92 to 1.49, but block-16 replacement increases it from 31.74 to 37.02, despite suppressing the local block-16 contrast. Thus, local suppression at the first amplification block does not establish lasting suppression. The clean token groups selected at Qwen blocks 16 and 18 are different.

### E.3 Formation and inheritance

For incoming residual $x$, attention update $a$, and MLP update $m$, direct norm gains are $g_{\text{attn}} = \|x + a\|_2 - \|x\|_2$ and $g_{\text{MLP}} = \|x + a + m\|_2 - \|x + a\|_2$. We average gains over selected tokens within each image and report the median across images. These describe the clean forward pass, not replacement effects.

Persistent onset requires at least half of images to contain a token above $10\times$ the median and a median selected-group variance share of at least .5 for three consecutive blocks. This unchanged rule

identifies Qwen block 16 and LLaVA block 10; it identifies no onset in Phi, InternVL, or Gemma on this sample. Qwen's MLP write rises with its tail ratio and variance share (Figure A3). At Phi block 13, median per-image direct attention/MLP gains on the top 2.5% are $-1.24/+117.79$. All 100 images contain strict tokens, but their median variance share is 23.1%, versus 57.2% for the top 2.5%. Phi therefore tests a broader tail. Gemma has no persistent high-norm group and its final-block change is analyzed separately.

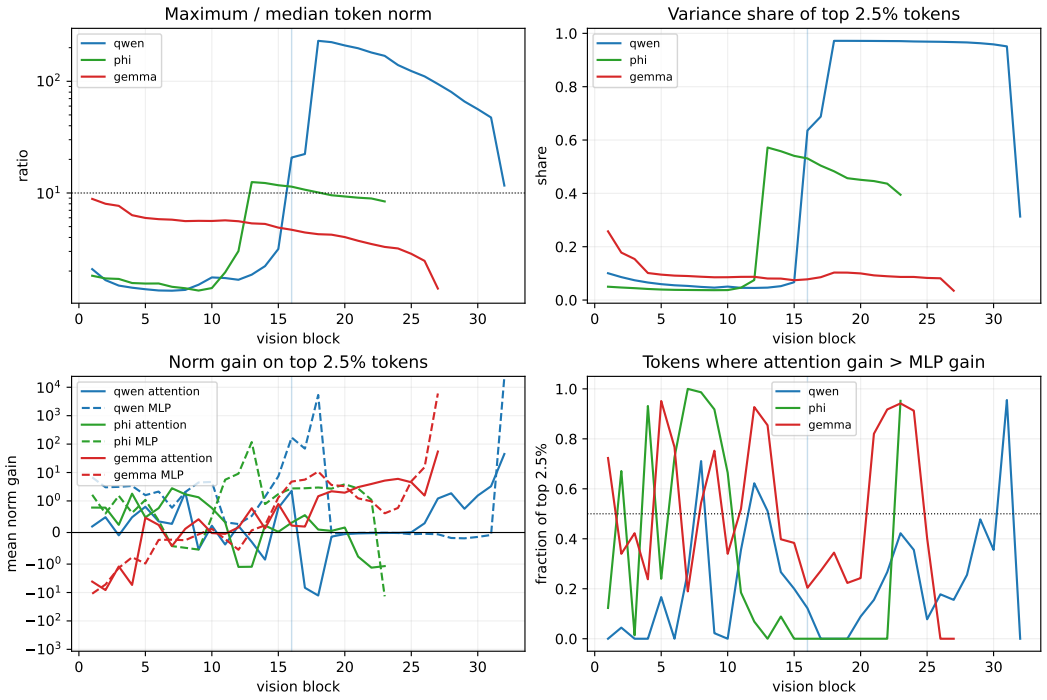



Figure A3: Depth-wise source attribution for Qwen, Phi, and Gemma on the 100-image SpatiaLab sample. Top left: maximum-to-median token norm ratio. Top right: share of total squared distance from the activation mean held by the top 2.5% of tokens. Bottom left: direct norm gains from attention (solid) and MLP (dashed) writes. Bottom right: the fraction of selected tokens for which attention gain is larger than MLP gain. Curves aggregate images by the median. Blue denotes Qwen, green Phi, and red Gemma throughout.

### E.4 LLaVA and InternVL source traces

LLaVA's operational onset is block 10. The mean paired attention-minus-MLP reduction in selected/control norm ratio has 95% CI $[-76.19, -74.54]$ there and $[-.185, -.177]$ at block 18: the later local effect is consistently small (Figure A4).

InternVL has strict tokens in 82% of images at block 12, with median variance share .099. At block 13, all images contain them; the median selected token fraction is .39% and median variance share is .545. The share is .523 at block 14 and .493 at block 15, so the three-block persistent-onset criterion is not met. We describe block 13 as an early amplification block. Its mean paired attention-minus-MLP norm reduction has 95% CI $[-8.93, -8.65]$. At block 18 the clean ratio remains 12.76, while both local replacements remove less than 1% of the ratio (median per-image fractional reduction). The propagated evidence is reported separately in Table A18.

### E.5 Internal MLP interventions

Figure A7 compares both tested depths for Qwen, Phi, and Gemma. Qwen's mean paired gate-minus-up difference in the fraction of norm ratio removed is $-.67$ percentage points at block 16 (95% CI $[-.82, -.50]$) and $+12.62$ at block 18 ($[11.83, 13.38]$). Both branches contribute strongly at block

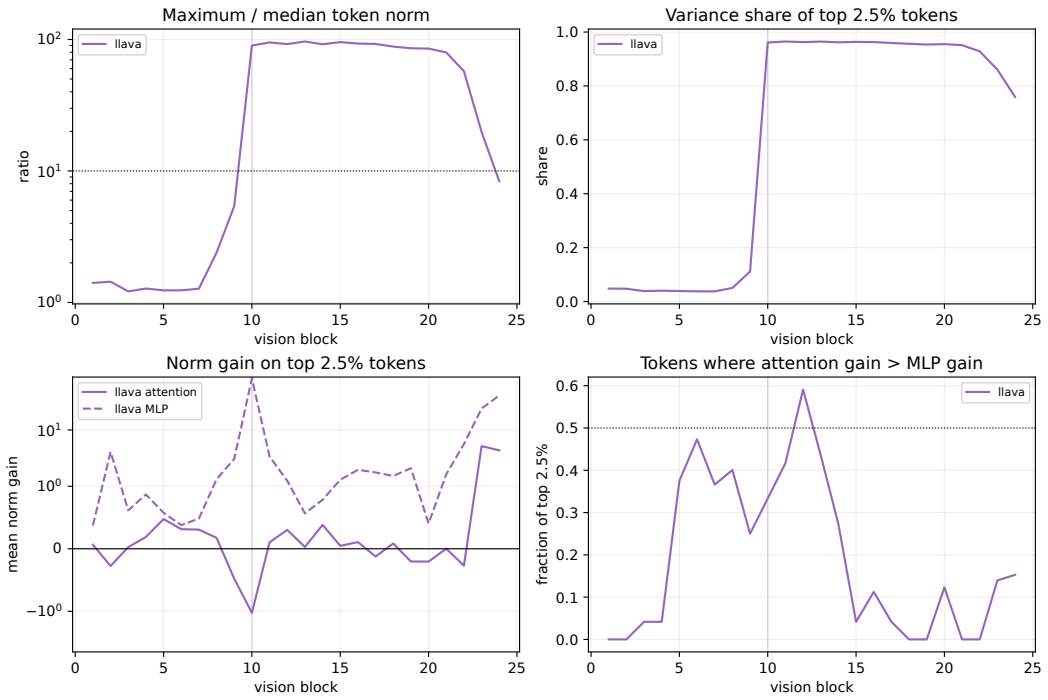



Figure A4: Depth-wise LLaVA source attribution on the 100-image SpatiaLab sample. The operational persistent onset is block 10. The attention and MLP curves are direct writes from the exact residual decomposition, not ablation effects.

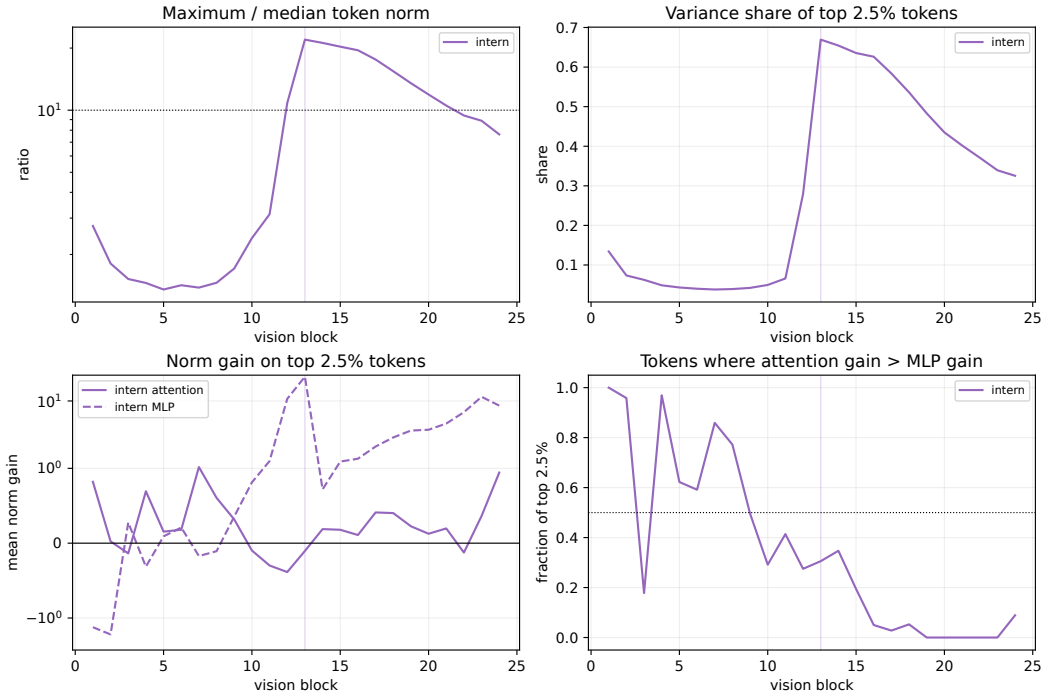



Figure A5: Depth-wise InternVL source attribution on the 100-image SpatiaLab sample. A precursor appears at block 12 and MLP amplification increases at block 13. The strict-group variance share exceeds .5 at blocks 13–14 but not for three consecutive blocks, so the persistent-onset rule is not met.

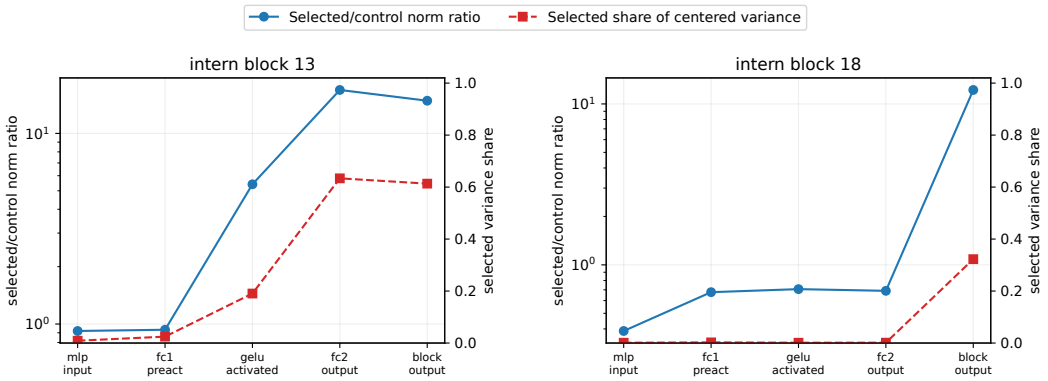



Figure A6: InternVL internal MLP attribution at blocks 13 and 18 on the 100-image SpatiaLab sample. Block 13 develops selectivity across GELU and FC2. Block 18 retains a large output contrast despite its small local MLP contribution.

16; the later contrast depends more on the gate branch. Phi's FC1-preactivation/post-QuickGELU replacements remove 73.8%/73.1% of the ratio at block 13 but only 1.0%/.6% at block 18 (medians of per-image fractional reductions).

For LLaVA, FC1-preactivation/post-GELU replacement removes 95.9%/96.0% of the ratio at block 10 but about .2% at block 18 (Figure A8). For InternVL block 13, selected/control ratios progress through input/FC1/GELU/FC2/output as $.93/.93/6.08/19.30/16.30$; selected variance share rises from .017 at FC1 to .151 after GELU and .599 after FC2. The two replacements remove 55.2%/55.2% at block 13 and .38%/.43% at block 18 (Figure A6). Gemma's FC1/post-GELU replacements remove 1.3%/1.2% at block 14 and 44.5%/42.1% at block 27, whose clean ratio is only 1.40. Gemma is a control without a comparable persistent concentrated tail, not a model without MLP effects. These tests establish contributions of the MLP paths, not sufficiency of GELU or QuickGELU alone.

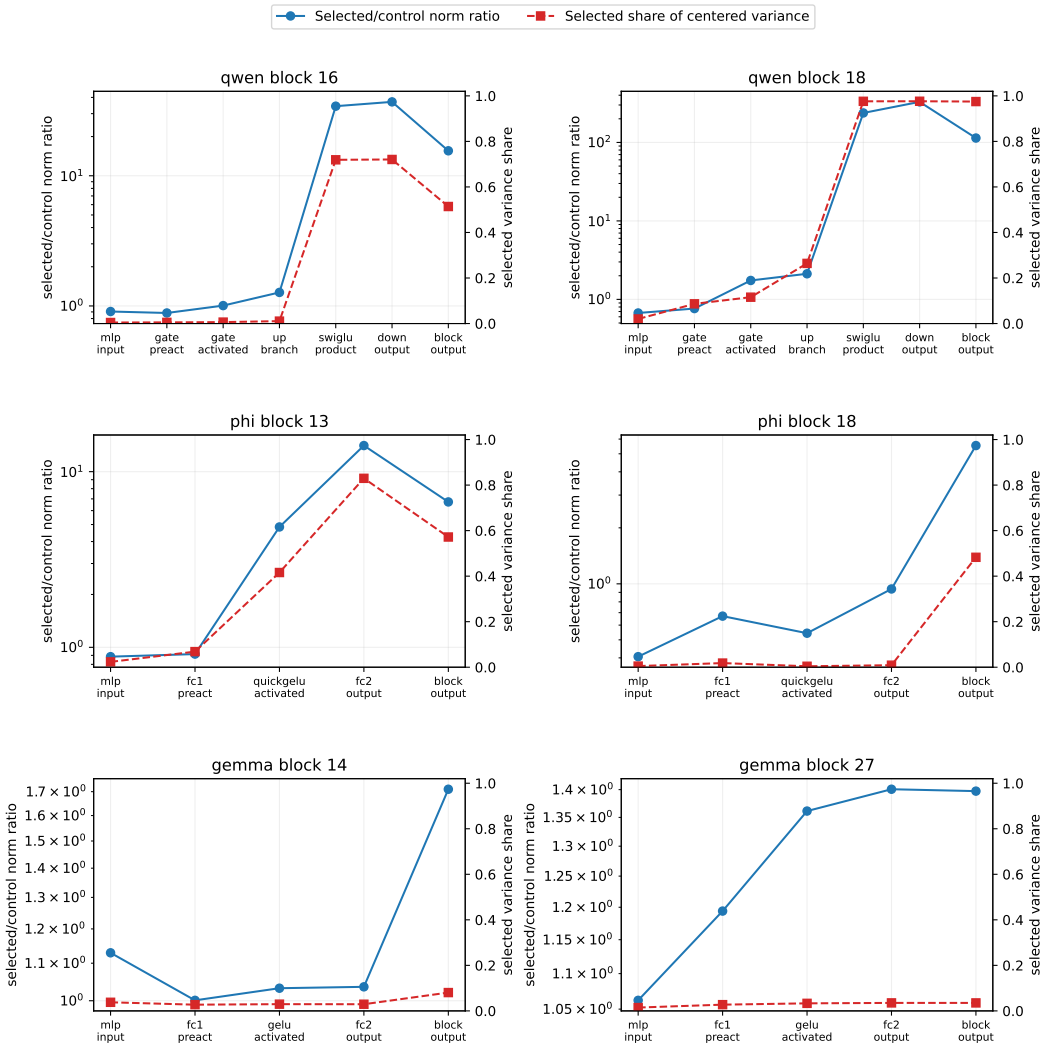



Figure A7: Full internal-MLP traces on the shared 100-image SpatiaLab sample (98 eligible images at Qwen block 16). Qwen blocks 16 and 18 use strict sink tokens; Phi blocks 13 and 18 and Gemma blocks 14 and 27 use the top 2.5%. The second Phi panel shows a large block-18 output contrast despite the small contrast in its local MLP output. Internal MLP inputs are normalized; these panels do not plot the incoming residual stream.

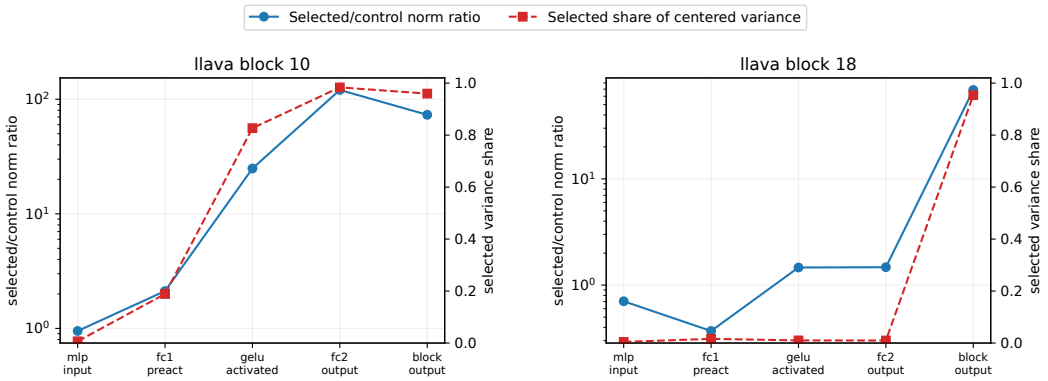



Figure A8: LLaVA internal MLP attribution at blocks 10 and 18 on the 100-image SpatiaLab sample. At block 10, contrast emerges across FC1, GELU, and FC2, and mean replacement tests their contributions. At block 18, the large output contrast persists while local MLP replacement has little effect.

## F GEMMA CONTROL AND FINAL-BLOCK ANALYSIS

Gemma's single dictionary uses width 4096, $\lambda = 1$, seed 49, and 20 training epochs on the 600-image CV-Bench Depth corpus. The image-level split contains 510 training and 90 validation images, with 1,024 tokens sampled per image using the same token indices across all blocks.

Among the 92,160 sampled validation tokens, none crosses $10\times$ the block median at any of 27 blocks. The cross-depth union mask therefore removes no tokens, and masked and raw aggregate EV are identical (0.86745). The 100-image SpatiaLab mechanism sample has occasional strict-threshold crossings without a persistent group or sharp onset (Appendix E).

Table A19 audits the single Gemma dictionary on 1,024 aligned tokens from each of 50 held-out images. At blocks 7/14/26, raw minus token-balanced EV is $.056/.058/-.030$. The top 2.5% holds 10.7%/9.6%/7.8% of centered variance, and removing it changes EV by .013/.018/.002. No token crosses the strict rule in this audit.

Table A19: Gemma held-out threshold-free EV. "Tok." gives every token equal weight, "Dec." averages ten equal-count norm groups, and "Typ." averages groups 1–9. "Top 2.5" is EV after removing the highest-norm 2.5%.

| Block | Raw | Tok. | Dec. | Typ. | d10 | Top 2.5 | Max/med. |
|---|---|---|---|---|---|---|---|
| 7 | .872 | .816 | .811 | .797 | .937 | .859 | 6.47 |
| 14 | .791 | .733 | .724 | .704 | .906 | .773 | 5.91 |
| 26 | .525 | .554 | .520 | .527 | .461 | .522 | 2.58 |
| 27† | .997 | .956 | .500 | .461 | .844 | .997 | 1.61 |

†Final-block control; excluded from the ordinary-depth comparison.

From blocks 1 to 26, the median norm grows 3.1-fold while max/median contracts (Table A20), unlike the expanding tails in the other encoders. This ratio is a useful diagnostic, but five towers cannot establish a universal cutoff. Block 27 has large norms across the population and max/median only 1.3 in the full-depth norm audit; it is analyzed separately.

A perfectly shared additive shift cancels under centering. Large absolute norms alone therefore do not explain pooled EV: concentration of centered variance and reconstruction errors must be checked. In the separate 50-image audit, block 27 has raw EV .997 but typical-token EV .461, so the absence of a strict high-norm tail does not imply agreement among all fidelity metrics. Gemma provides a fidelity control with one dictionary, not a dictionary-ranking or behavioral restoration experiment.

Table A20: Gemma-3 SigLIP token-norm quantiles. Block 27 shows a separate population-wide increase in token norms.

| Block | Median | p99 | p99.9 | Max | Max/median |
|---|---|---|---|---|---|
| 1 | 25.4 | 75.8 | 184.5 | 248.2 | 9.8 |
| 4 | 23.8 | 57.0 | 99.4 | 171.8 | 7.2 |
| 8 | 24.8 | 57.4 | 82.9 | 157.8 | 6.4 |
| 12 | 25.4 | 55.8 | 87.5 | 160.5 | 6.3 |
| 16 | 32.9 | 57.2 | 92.2 | 165.8 | 5.0 |
| 20 | 45.4 | 69.6 | 170.0 | 205.3 | 4.5 |
| 24 | 59.5 | 92.4 | 179.8 | 215.6 | 3.6 |
| 26 | 79.7 | 125.9 | 182.4 | 217.7 | 2.7 |
| 27 | 5360.4 | 6789.4 | 6931.3 | 7065.5 | 1.3 |

## G NORM-GROUP VISUALIZATIONS ACROSS ENCODERS

Figures A9–A13 show separate full-width t-SNE visualizations (van der Maaten & Hinton, 2008) of SpatiaLab token activations for Qwen, LLaVA, InternVL, Phi, and Gemma. Each figure retains all three evaluated depths and both representations: raw activations in the top row and unit-normalized activations in the bottom row. Colors identify the original activation-norm deciles, with 10 the highest. Normalization removes vector magnitude before projection, allowing inspection of directional structure associated with the norm groups.

For each model and block, we sample 150 tokens per norm decile from activations across the 1,400 SpatiaLab images. Raw and unit-normalized views use the same tokens. We reduce each representation to 50 principal components before applying t-SNE with perplexity 30 and 1000 iterations.

Qwen, LLaVA, and InternVL develop more visible norm-group structure at later depths, including in the normalized views. Phi also shows a localized highest-norm subgroup at blocks 12 and 18 that remains visible after normalization, while most norm groups overlap. Gemma instead shows broader gradients and overlap. These patterns complement the fidelity checks in Appendix D; a two-dimensional projection does not establish a discrete feature or behavioral importance. Distances across separately fitted projections are not directly comparable. The token restoration and mechanism interventions in Appendices A.4 and E test behavioral contributions and the necessity of model updates, respectively.

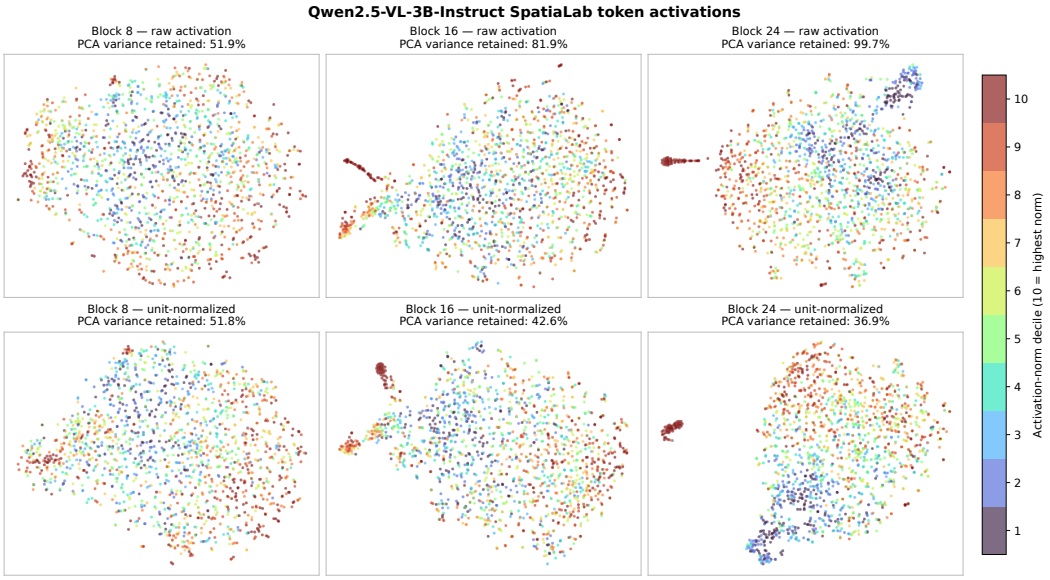



Figure A9: **Qwen2.5-VL norm-group structure on SpatiaLab.** Columns show blocks 8, 16, and 24; rows show raw (top) and unit-normalized (bottom) activations. Colors denote original activation-norm deciles. A highest-norm subgroup becomes distinct at blocks 16 and 24 and remains visible after normalization, consistent with directional as well as magnitude differences. Most other tokens continue to overlap.

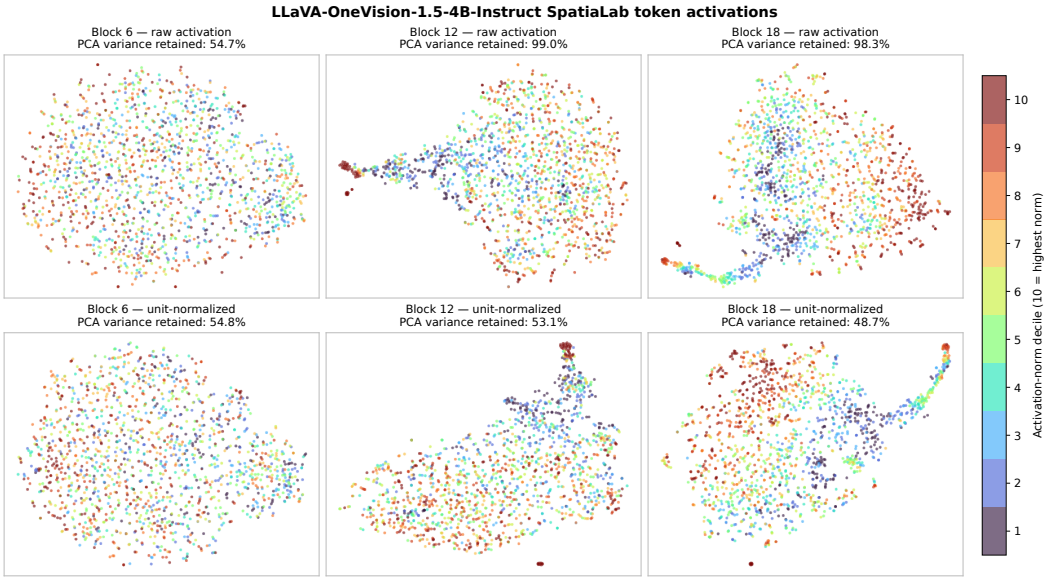



Figure A10: **LLaVA-OneVision norm-group structure on SpatiaLab.** Columns show blocks 6, 12, and 18; rows show raw (top) and unit-normalized (bottom) activations. Colors denote original activation-norm deciles. Groups overlap broadly at block 6 and develop more visible structure at blocks 12 and 18. Structure persists after normalization, although the norm groups do not form ten separate clusters.

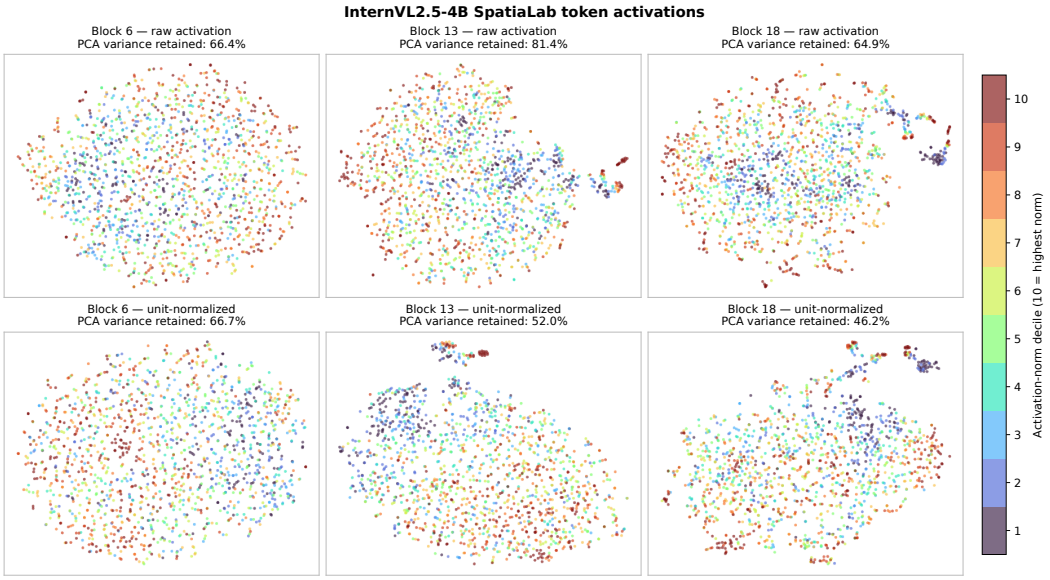



Figure A11: **InternVL2.5 norm-group structure on SpatiaLab.** Columns show blocks 6, 13, and 18; rows show raw (top) and unit-normalized (bottom) activations. Colors denote original activation-norm deciles. Groups overlap at block 6; later views contain smaller separated groups alongside a broad mixed population. Their persistence after normalization shows structure associated with direction, while substantial overlap remains.

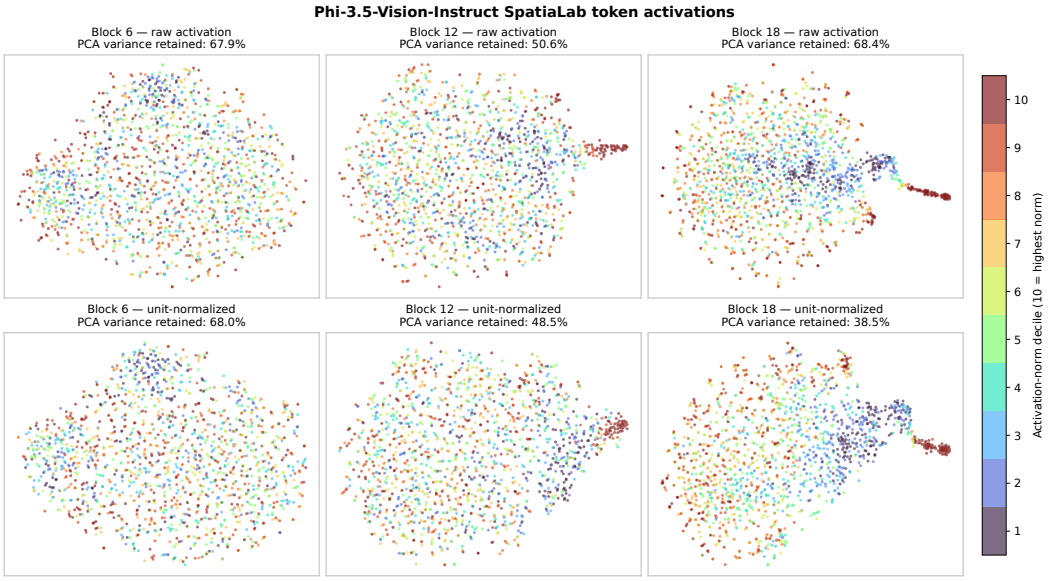



Figure A12: **Phi-3.5-Vision norm-group structure on SpatiaLab.** Columns show blocks 6, 12, and 18; rows show raw (top) and unit-normalized (bottom) activations. Colors denote original activation-norm deciles. Groups overlap broadly at block 6. At blocks 12 and 18, a localized highest-norm subgroup is visible in both representations, indicating directional as well as magnitude differences. Most tokens remain in a mixed population.

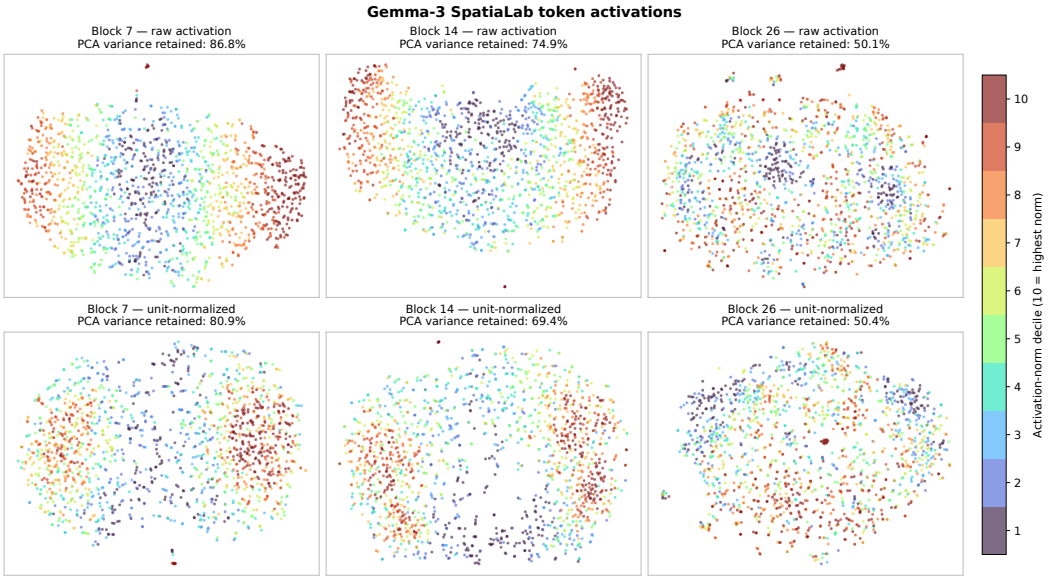



Figure A13: **Gemma-3 norm-group structure on SpatiaLab.** Columns show blocks 7, 14, and 26; rows show raw (top) and unit-normalized (bottom) activations. Colors denote original activation-norm deciles. The views show broad gradients and overlap without a comparably detached highest-norm group. Block 27 is excluded because its final-block shift is examined separately in Appendix F.