ROGUE DIMENSIONS MASK A LEARNED CORTICAL TRAJECTORY OF VISUAL TOKENS IN VISION–LANGUAGE DECODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) process image tokens with an entire language decoder, yet brain-alignment studies usually score the visual encoder, one layer, or the final hidden state. We trace the representational geometry of image-token positions after every decoder block of 13 open checkpoints (256M–10B parameters) and compare it with seven visual-cortex regions in eight Natural Scenes Dataset (NSD) participants and, for 10 checkpoints, nine participants of the independent Natural Object Dataset (NOD). Two depth patterns hold under the frozen correlation-distance protocol and under three feature-scaling controls, and are absent from depth-matched untrained decoders: alignment with the lateral stream rises over the first three quarters of the decoder (+0.040 to +0.046 of the noise ceiling), and early-cortex alignment falls in the last quarter (negative in 13 of 13 NSD checkpoints; participant-by-stimulus intervals exclude zero in 7), as does alignment with the other streams. The large change at the returned final state, which at image positions no later computation reads, is instead a measurement effect: three channels carry a median 83% of the energy of the mean image token at the last read state and 34% after the output normalization, and standardizing features shrinks the final lateral gain from +0.030 to +0.005. Feature scaling also decides which checkpoints appear to move from early toward lateral geometry (6 of 13 raw, 11 standardized). Endpoint directions agree across datasets in 19 of 20 cells under the frozen metric and in 14 with standardized features, where the lateral rise transfers and the early decline is uncertain on NOD. The depth decomposition was defined after inspecting the curves and is exploratory. Layer-wise VLM–brain comparisons should standardize features and report the last read state alongside final-layer scores.

Keywords: representational similarity analysis; vision–language models; Natural Scenes Dataset; visual cortex; lateral visual pathway; brain alignment

1 INTRODUCTION

Modern vision–language models (VLMs) do not stop at a visual encoder. They project an image into a sequence of visual tokens and pass those tokens through every block of a language model before producing text. This long decoder trajectory is important: model-side analyses show that visual tokens change their semantic accessibility, sparsity, dimensionality, and contribution to prediction with depth (Neo et al., 2025; Liu et al., 2025; Venhoff et al., 2025; Fan et al., 2026; Ghosh et al., 2026). Yet neural comparisons usually evaluate a vision encoder or one selected hidden state. Consequently, strong brain alignment at the visual interface does not tell us whether the geometry survives, strengthens, or is reorganized inside the language stack.

The missing comparison is not simply "which layer best predicts cortex?" Human visual cortex contains distinct representational geometries: early areas emphasize local visual structure, whereas ventral, lateral, and parietal territories differ in their sensitivity to objects, actions, and spatial structure. A decoder could therefore lose alignment everywhere, preserve its input geometry, or become closer to some cortical territories while moving away from others. A single final-layer score can-

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.

not distinguish these possibilities. Nor can one architecture establish that a trajectory is a general property of VLMs.

We measure the geometry of the image-token positions at the decoder input and after every language block. For each readout, we compute a stimulus representational dissimilarity matrix (RDM) and compare it with region-specific fMRI RDMs. The primary dataset is the Natural Scenes Dataset (NSD; eight participants and 515 equally repeated images) (Allen et al., 2022); an independent boundary test uses the repeated-image COCO experiment of the Natural Object Dataset (NOD; nine participants and 120 images) (Gong et al., 2023). Our checkpoint panel was fixed before the corrected NSD brain preprocessing was complete and contains 13 open VLM checkpoints spanning five decoder lineages and 256M–10B parameters. We treat these checkpoints as a fixed panel rather than a random sample from a population of all VLMs.

The study makes four contributions. First, it supplies a measurement protocol for complete decoder trajectories, with explicit image-position masks, multiple pooling rules, precision checks, and cryptographic provenance; participants, not model–subject cells, are the unit of inference, and an independent dataset serves as a boundary test. Second, it shows that raw correlation-distance RSA of decoder states is dominated late in the stack by a few high-magnitude channels, which the output normalization then suppresses: the resulting jump in alignment occurs in a state that, at image positions, no later computation reads, and it largely disappears when features are standardized. Third, it identifies two depth patterns that survive this and two further metric controls and that untrained decoders of the same depth do not show: a rise of lateral-stream alignment over the first three quarters of the decoder and a fall of early-cortex alignment, shared by the other streams, in the last quarter. Fourth, it shows that which checkpoints look brain-like at the endpoint depends on feature scaling, so checkpoint-level contrasts need a scaling control. The depth decomposition was defined after the curves had been inspected and is reported as exploratory throughout.

2 Related work

Models and visual cortex. Representational similarity analysis (RSA) provides a common geometry in which brains and models can be compared without fitting a high-capacity readout (Kriegeskorte et al., 2008; Nili et al., 2014). Work on NSD has shown that both visual and language-derived representations align with high-level visual cortex (Doerig et al., 2025; Ponnambalam & Pottore Venkiteswaran, 2026). Oota et al. (2025) fit layer-wise encoding models from the text-response embeddings of instruction-tuned multimodal models to NSD responses in early and category-selective regions; in two of their three models, later layers predict early visual regions best. These results motivate a regional, depth-resolved analysis, but they do not follow the image-token geometry from the visual interface to the decoder output.

The closest brain-alignment studies narrow the novelty of our claim. Bavaresco et al. perform layer-wise RSA for several VLMs against fMRI concept representations (Bavaresco et al., 2026), Oota et al. (2025) resolve both layers and regions on NSD, and Shah et al. (2026) compare 12 open VLMs and eight NSD participants. Layer-wise VLM–brain comparison and broad NSD checkpoint panels are therefore not new. Our distinct object of measurement is the sequence of hidden states at the *image-token positions* through the complete language decoder, decomposed by depth, compared with seven cortical streams, and tested on an independent fMRI dataset.

Visual tokens inside language models. Mechanistic work has localized object information in visual tokens and traced their changing vocabulary interpretability across LLaVA layers (Neo et al., 2025). Subsequent studies analyze the evolution of visual information, layer-wise token semantics, representation compression, and sparse or redundant token classes across broader VLM panels (Liu et al., 2025; Krojer et al., 2026; Fan et al., 2026; Ghosh et al., 2026), and how visual features map into language feature space in a controlled model that joins a frozen encoder and a frozen language model with a trained adapter (Venhoff et al., 2025). Attention knockout shows that text reads image information mainly in middle layers, with early and late layers contributing little (Kaduri et al., 2025), and late visual states become largely interchangeable across depth (Ghosh et al., 2026). These model-side targets answer what tokens encode or whether they are useful. Cortical RDMs provide a different reference geometry: they ask whether stimulus relations in a hidden state resemble relations measured in a particular human visual territory. We do not interpret similarity as mechanistic identity

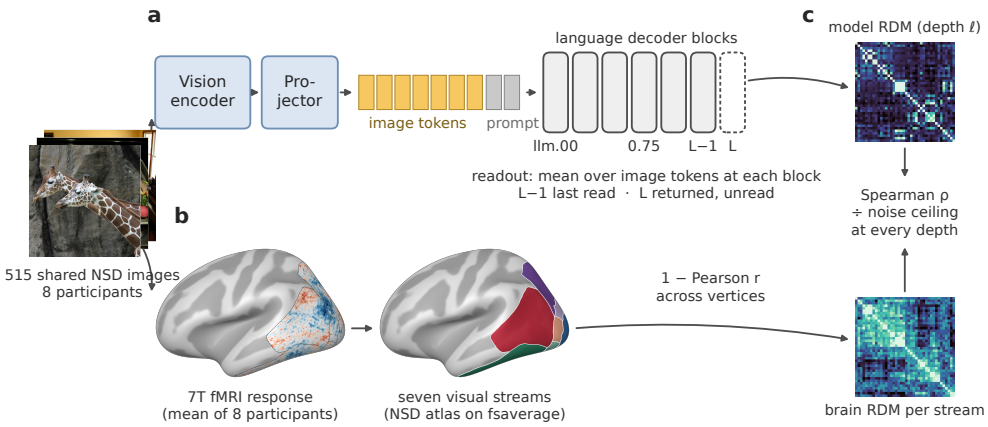

Figure 1: The comparison, end to end. **(a)** A vision–language model turns one shared image into visual tokens that pass through every block of its language decoder; at each block we pool the image-token states into one representation. The input to the last block, $L - 1$, is the last image-token state that generation reads, whereas the returned state L feeds no later computation at image positions. **(b)** NSD participants saw the same images in 7T fMRI; responses in the seven `streams` regions give one RDM per participant and region (shown: the group-mean response to this image, and the atlas on the inflated surface). **(c)** Model and brain RDMs are compared with Spearman correlation, scaled by the lower noise ceiling, at every depth; that ratio is what the later figures plot. Photographs: COCO (Lin et al., 2014), CC BY 2.0.

or biological plausibility. Recent critiques show that model–brain conclusions can change with the similarity measure (Soni et al., 2024) and that standard alignment scores can fail to discriminate between models or to reproduce a region's characteristic pattern of cross-region alignment (Höfling et al., 2026). This is why our claims are comparative within a fixed RSA protocol, report raw and ceiling-scaled values, and emphasize failures of generalization alongside positive effects.

3 Methods

Figure 1 summarizes the measurement: the readouts taken from a decoder, the cortical targets they are compared with, and the quantity every later figure reports.

3.1 Datasets and cortical targets

NSD. We use the eight NSD participants and the GLMsingle-denoised `betas_fithrf_GLMdenoise_RR` (b3) single-trial estimates in fsaverage surface space (Allen et al., 2022; Prince et al., 2022). The preprocessing order is load-bearing. For every scan session and vertex, we estimate a mean and population standard deviation from all 750 trials, standardize those trials, and only then select shared-image presentations. An earlier version of our pipeline standardized after selecting the shared trials; all reported results were recomputed after correcting this error. We retain the 515 images presented at least three times to every participant and average exactly three repeats per image. Vertices that are non-finite in any participant are removed by a common mask.

We use the seven bilateral regions in NSD's `streams` atlas: early, mid-ventral, ventral, mid-lateral, lateral, mid-parietal, and parietal. The primary regions, fixed before corrected results were inspected, are early and lateral cortex. The remaining five provide secondary anatomical context.

NOD boundary test. The NOD COCO experiment presents 120 natural images ten times to nine participants (Gong et al., 2023). We use the released surface GLM beta maps in fsLR32k space, select the balanced published runs 1–10 for every participant, and verify image names exactly across all 90 run files. Within every run, each grayordinate is standardized across all 120 image estimates

before repetitions are averaged; this prevents arbitrary run-level beta scale from changing its weight. We transfer the NSD streams atlas from fsaverage to fsLR using registered spherical coordinates and nearest-neighbour assignment. Median transfer error is approximately $0.20°$ and every ROI's share of the labelled visual territory differs from the source by less than 5% (the prespecified gate was 10%). Odd/even-run reliability is computed across images separately at each grayordinate and Spearman–Brown corrected. These checks validate the transfer but do not make NSD and NOD identical acquisitions; NOD is used as an external boundary test.

3.2 Checkpoint panel and decoder readouts

The panel contains 13 publicly available checkpoints: three SmolVLM sizes; Qwen2-VL-2B/7B and Qwen2.5-VL-3B; LLaVA-OneVision-0.5B; InternVL3-2B/8B; PaliGemma2-3B-pt, 3B-mix, and 10B-pt; and Idefics2-8B. They span five decoder lineages (SmolLM2, Qwen2, Qwen2.5, Gemma2, and Mistral). We use the public releases described by their model reports (Marafioti et al., 2025; Wang et al., 2024; Bai et al., 2025; Li et al., 2025; Zhu et al., 2025; Steiner et al., 2024; Laurençon et al., 2024). LLaVA-1.6-Mistral-7B was attempted but excluded before analysis because its expanded image positions could not be validated under the common extraction protocol. Checkpoints are therefore a deliberately chosen, fixed panel; model-level summaries are descriptive and do not estimate a population effect over all VLMs. Twelve checkpoints use causal attention throughout. PaliGemma2 is a prefix language model: attention is bidirectional within the prefix, which here holds the image tokens and the empty prompt, so its image tokens also attend to later image positions. The argument of Section 4.3 that the returned final state is unread holds for both masking regimes, because no block follows the last one.

Each model receives one image and the minimal model-native empty-prompt template. Native processors are used with image tiling disabled so that a single image is not silently expanded into a model-specific number of crops. Activations are extracted in float32. The processor/model-declared image token identifier defines a Boolean mask, and extraction fails unless the expected image positions are present. For InternVL3-8B, the checkpoint that produced the most unusual trajectory, we additionally verified on the extraction machine that all 256 image tokens form the expected contiguous span, non-image positions are invariant to the input image before attention, and 28 decoder blocks produce 29 hidden states.

Let $H_{m\ell it} \in \mathbb{R}^d$ denote the hidden vector of checkpoint m, readout ℓ, image i, and image-token position t. The primary representation is

$$z_{m\ell i} = |T_i|^{-1} \sum_{t \in T_i} H_{m\ell it}. \tag{1}$$

Readout `llm.00` is the embedding-side state after visual embeddings have been inserted but before the first decoder block; readout L is the model's returned final hidden state, after the last block and its output normalization. The cached interface does not expose the final block before that normalization, so terminal effects cannot be attributed uniquely to one of those operations. At image positions, readout L also feeds no later computation, whereas readout $L-1$ is the last state that generation reads (Section 4.3). Sensitivity analyses use an elementwise token median and a trimmed mean that drops the top 1% of image tokens by norm. We also start at `llm.01` and omit the last 10% or 25% of readouts to detect endpoint-driven curves. Reduced-precision and generation-position measurements are controls, not part of the primary estimand.

3.3 Representational similarity and scaling

For every model readout and every participant–ROI pair we form a condensed RDM over the same ordered images using correlation distance, $1 - r$, across features or vertices. A model–brain score is Spearman's correlation between their RDM vectors. Calculations use float64 accumulation and exact rank caching; 515 NSD images yield 132,355 pairwise distances per RDM.

For each ROI, the lower RSA noise ceiling is the mean Spearman correlation between each participant's RDM and the mean of the other participants' rank-transformed RDMs; the upper ceiling includes that participant in the group mean (Nili et al., 2014). Pooling ranks rather than raw distances, as rsatoolbox does, keeps any participant's distance scale from setting its weight in the reference.

An earlier version of this analysis pooled raw distances; the change moves ceiling-scaled values by at most about 5% and alters no sign or test. Scores are divided by the lower bound so that regions of different reliability share one scale; raw RSA and upper-bound scaling are stored in the artifact. A ceiling-normalized correlation is a fraction of an RSA ceiling, not explained variance, and division by a positive, layer-invariant ceiling leaves every trajectory sign unchanged. As a control added after the frozen analysis, we also z-score every model feature across images before forming its RDM, which mirrors the per-vertex standardization of the fMRI data (Section 4.4).

3.4 Frozen estimands and inference

The primary regions, endpoint, pooling and tests were frozen before the corrected NSD matrix was computed, but after legacy results and an approximate diagnostic had been inspected, so the specification is not a preregistration. A second version, prompted by an independent audit, changed only the ceiling pooling and the number of joint bootstrap draws, and added the exploratory depth decomposition of Section 4.3.

The primary checkpoint-level estimand for participant s is the endpoint change

$$\Delta_{msr} = a_{msrL} - a_{msr0}, \tag{2}$$

where a is lower-ceiling-scaled RSA and $r \in \{\text{early}, \text{lateral}\}$. We report the mean Δ, its percentile 95% interval from 10,000 participant bootstrap draws, the number of positive participants, and a two-sided Wilcoxon signed-rank test. The 26 NSD checkpoint-by-primary-ROI tests form one Holm family. Secondary curve descriptors are an ordinary least-squares slope over relative depth and Spearman correlation with depth.

As a generalization check over both sampled people and sampled images, we additionally draw 10,000 paired participant-by-stimulus bootstrap samples for the endpoint and for the depth steps of Section 4.3. For each draw, the same participant and image indices are used for every checkpoint, estimand and ROI. We rebuild the ranked RDM comparison after sampling; pairs in which two sampled positions refer to the same original image are excluded rather than treated as genuine zero-distance observations. The full-sample lower ceiling is held fixed, so the interval reflects numerator uncertainty. This is an ordinary crossed percentile bootstrap, not the variance-corrected procedure of Schütt et al. (2023); with eight or nine participants its calibration is limited, so we read it as a robustness description rather than a test. All draws are saved with the result. An earlier version used 1,000 draws, too few for a stable bound on the NOD lineage-balanced lateral interval.

To summarize the fixed panel without pseudo-replication, we first average checkpoint effects within each participant. We show both a checkpoint-weighted summary and a lineage-balanced summary (checkpoints averaged within decoder lineage, then lineages equally weighted), with leave-one-lineage-out values. These subject-level summaries use the same bootstrap and signed-rank test, with Holm correction across the two primary ROIs. Cross-dataset transfer is reported cell by cell for checkpoints present in both datasets: a directional match requires equal signs, while a strict match additionally requires both participant-bootstrap intervals to exclude zero. NOD lacks three large checkpoint readouts; they remain explicit missing cells rather than being imputed.

The analysis specification, exact input hashes, software versions, and checkpoint names are emitted with the result artifact. The representation caches record exact derived values, although the original weight and processor revisions were not saved at extraction time; this limits bit-identical re-extraction from future moving repository heads.

4 Results

4.1 Decoder trajectories are prevalent but not universal on NSD

Every checkpoint's endpoint contrast is listed in Table 1 and plotted in Figure 8. Early-cortex alignment decreases from `llm.00` to the final decoder state in 9 of 13 checkpoints, and lateral alignment increases in 10 of 13. The two directions co-occur in only 6 checkpoints; in the others both regions move the same way. The other higher-order streams move with the lateral stream, most closely the ventral one (Table 2), so the endpoint contrast is between early and higher-order cortex rather

than specific to the lateral stream; which checkpoints show it also depends on feature scaling (Section 4.4). Per-checkpoint intervals, all seven regions, and lineage summaries are in Appendix A (Figures 10 to 12 and table 2).

The checkpoint-weighted participant summary is -0.028 [-0.034, -0.023] in early cortex and +0.030 [+0.022, +0.036] laterally (fractions of the lower RSA ceiling). Equal weighting of decoder lineages yields -0.020 [-0.024, -0.015] and +0.033 [+0.027, +0.039], respectively. We report these summaries to show how the chosen panel behaves, not as model-population tests. The corresponding participant-by-stimulus intervals are [-0.0418, -0.0152] and [+0.0104, +0.0485] for checkpoint weighting, and [-0.0332, -0.0070] and [+0.0141, +0.0519] for lineage weighting. Checkpoint-level p-values are adjusted jointly over all 26 primary cells; none survives that correction (minimum adjusted $p = 0.2031$), as the exact signed-rank test with eight participants cannot reach the first Holm threshold. Both panel summaries survive adjustment over their two primary ROIs ($p_{\mathrm{Holm}} = 0.0156$).

4.2 NOD IS A BOUNDARY TEST, NOT A BLANKET REPLICATION

NOD provides 9 new participants and a different acquisition, surface, and 120-image set. Of the 20 early/lateral cells available in both datasets, 19 retain their sign (Figure 13); 17 also have participant-bootstrap intervals excluding zero in both datasets, a criterion that conditions on each dataset's images. The directions therefore largely transfer, but this is support for checkpoint-specific directions, not a universal law. NOD's checkpoint-weighted effects are -0.039 [-0.046, -0.032] for early cortex and +0.077 [+0.051, +0.107] for lateral cortex; the lineage-balanced counterparts are -0.044 [-0.050, -0.039] and +0.069 [+0.043, +0.099]. Their participant-by-stimulus intervals are [-0.1067, +0.0244] and [+0.0043, +0.1526] under checkpoint weighting, and [-0.1182, +0.0264] and [-0.0053, +0.1473] under lineage weighting. Of these four intervals only the checkpoint-weighted lateral one excludes zero; both early intervals and the lineage-balanced lateral interval include it, reflecting the larger stimulus uncertainty of the 120-image set. Three large checkpoints lack NOD representations and remain visible as dashes; no value is imputed and no reduced-precision local extraction is substituted.

4.3 WHERE ALONG THE DECODER THE CHANGE HAPPENS

The endpoint compresses each curve into two points, but the curves in Figure 2 are not monotone, and one of the two points has a special status. Readout $L - 1$, the input to the final decoder block, is the last image-token state that generation reads: the final block's attention turns it into the keys and values that later text positions attend to. Readout L, the final block's output at image positions plus the output normalization, is read by nothing, because no block follows it and next-token prediction reads only the final text position. After inspecting the curves we therefore decomposed each endpoint into three additive steps, from `llm.00` to 0.75 relative depth, from there to $L - 1$, and the terminal step from $L - 1$ to L (Figure 3). The decomposition is exploratory but uses the same participant-level and joint participant–stimulus inference as the primary endpoint.

The three steps behave differently, with the same directions in both datasets. Over the first three quarters, lateral alignment rises (NSD +0.040, joint interval [+0.0273, +0.0528]; NOD +0.064) while early alignment changes little on average (-0.003 on NSD). Between 0.75 depth and $L - 1$, early alignment falls in 13 of 13 NSD checkpoints (participant$\times$stimulus intervals excluding zero in 7) and 9 of 10 NOD checkpoints (NSD -0.032 [-0.0409, -0.0233]; NOD -0.033, with a joint interval that includes zero), although by very different amounts; about 89% of the NSD panel's early decline accumulates after 0.75 depth. Lateral alignment falls there too, on NSD by more (-0.041 [-0.0530, -0.0291], negative in 11 of 13 checkpoints), and so does the ventral stream (Section 4.4). The terminal step then raises lateral alignment in 11 of 13 NSD checkpoints (+0.030 [+0.0203, +0.0404]).

Measured at the last state that generation reads, the NSD panel's lateral gain disappears (-0.001 [-0.0163, +0.0140]), the early decline persists (-0.035 [-0.0470, -0.0234]), and 3 rather than 6 checkpoints show both directions. On NOD the last-read lateral change stays positive (+0.029), but its joint interval [-0.0243, +0.0840] includes zero. Under the frozen metric the returned state thus supplies the NSD panel's lateral gain; the next subsection shows how much of that is metric.

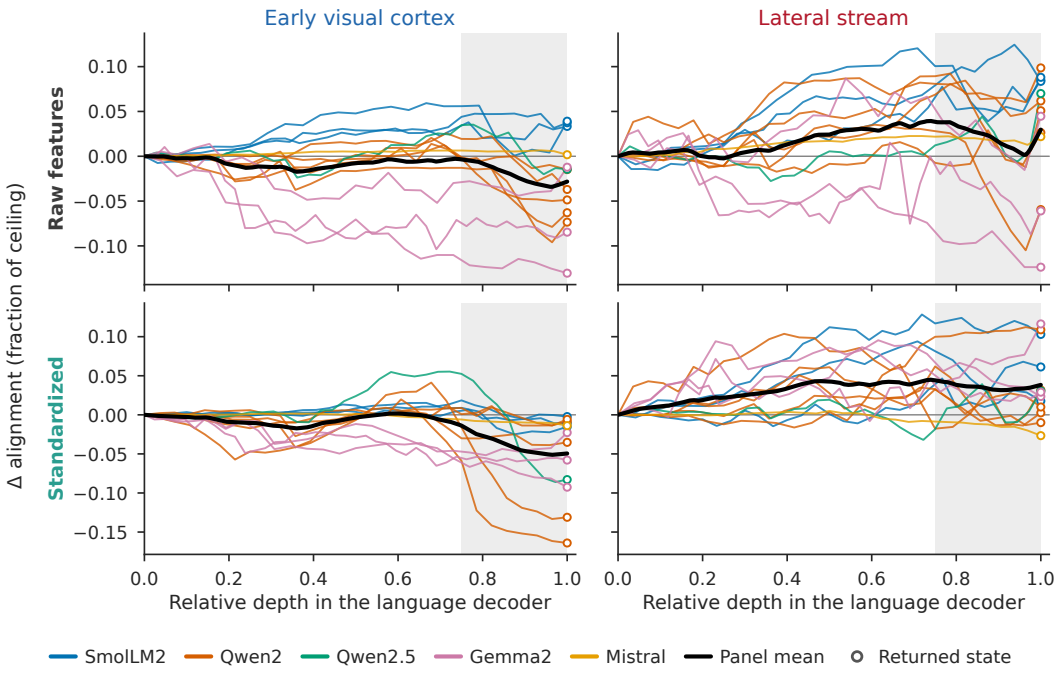

Figure 2: Change in alignment from the decoder input (fraction of the lower RSA noise ceiling) across relative depth for the thirteen NSD checkpoints (thin lines, coloured by decoder lineage) and their mean (black), in early visual cortex (left) and the lateral stream (right). **Top:** the frozen protocol on raw features. **Bottom:** features standardized across images (Section 4.4). Hollow points mark the returned final state, which at image positions no later computation reads; the last quarter of depth is shaded.

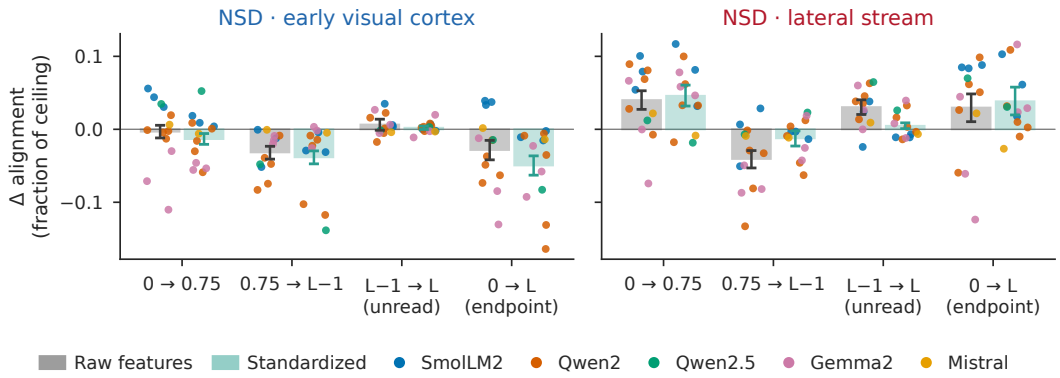

Figure 3: The endpoint decomposed by depth on NSD, for raw (grey) and standardized (teal) features. Bars: checkpoint-weighted panel change as a fraction of the lower RSA ceiling with 95% participant×stimulus intervals (10,000 draws); dots: individual checkpoints, coloured by decoder lineage. The first three steps add up to the fourth; the terminal step changes a state that no later computation reads. The decomposition was defined after inspecting the depth curves; NOD is in Figure 9.

4.4 Rogue dimensions and the returned final state

Correlation distance across features lets high-magnitude channels dominate, and transformer residual streams develop such channels (Timkey & van Schijndel, 2021; Sun et al., 2024). In our readouts the three largest channels carry a median 23% of the energy of the mean image token at `llm.00`, 64% at 0.75 depth, 83% at $L-1$, and 34% after the output normalization, which only the returned state passes through and which rescales channels with learned gains. We therefore recomputed every NSD trajectory with each feature standardized across images before forming its RDM, and with two further controls: removing the three highest-energy channels, and ranking each image's features.

None of these metrics is part of the frozen protocol; the values below are checkpoint-weighted participant means.

Two depth patterns survive all four metrics (Figure 6). Lateral alignment rises over the first three quarters (+0.040 raw, +0.046 standardized, +0.045 without the top channels, +0.045 on ranks), and early alignment falls in the last quarter (-0.032, -0.038, -0.019, -0.024). The terminal lateral gain does not survive: it shrinks from +0.030 to +0.005 under standardization and to +0.011 when only three channels are removed. The late declines of the ventral and lateral streams shrink as well (lateral -0.041 to -0.012, ventral -0.049 to -0.029), whereas standardization leaves the early decline intact and the parietal declines large. Most of the terminal rebound, and part of the late dip, therefore reflect how the last read and returned states are scaled rather than their stimulus geometry. Scaling also changes which checkpoints show the endpoint pattern: early alignment falls and lateral alignment rises in 6 of 13 checkpoints under the raw metric but in 11 after standardization, where the SmolVLM and PaliGemma2 exceptions disappear.

Untrained decoders do not produce the two robust patterns. For the 9 checkpoints with depth-matched, randomly initialized language stacks behind the same vision tower and projector (2–3 seeds each), lateral alignment falls over the first three quarters (-0.029 raw, -0.044 standardized) where the trained stacks raise it (+0.045, +0.051), and early alignment barely changes in the last quarter (-0.003, against -0.039 for the trained stacks). Both patterns therefore depend on trained decoder weights, and the untrained stacks show no terminal jump.

Rerunning the complete analysis on standardized features, with the same participant-by-stimulus bootstrap, confirms both patterns on NSD: lateral alignment rises by +0.046 [+0.0319, +0.0606] up to 0.75 depth, early alignment falls by -0.038 [-0.0474, -0.0296] from there to $L - 1$, and the endpoint is more consistent across checkpoints than the raw one (early negative in 13 of 13, lateral positive in 11, terminal lateral step +0.005; Figure 3, cortical maps in Figure 5). On NOD the lateral rise transfers (endpoint positive in 9 of 10 checkpoints) but the early decline does not (negative in 6, joint interval [-0.1161, +0.0128]; Figure 9), so endpoint directions agree across datasets in 14 of 20 cells, against 19 under the raw metric.

4.5 Sensitivity and a terminal-state counterexample

Primary signs are compared under trimmed-mean and median token pooling, a start at `llm.01`, lower- versus upper-ceiling scaling, raw RSA, and removal of the final 10% or 25% of states; every variant is in the hashed artifact, so none was selected for agreeing with the primary analysis. The most fragile case, InternVL3-8B, keeps its negative lateral endpoint under every pooling rule and precision, but it turns positive once the last 10% of states are removed, on both datasets (Appendix A.6).

5 Discussion

Following visual tokens through a language decoder turns "is this VLM brain-aligned?" into three questions: which cortical geometry each internal state resembles, whether the measurement is dominated by a few channels, and whether the model uses the state. Across the fixed panel, trained decoders make image tokens more similar to higher-order visual cortex over most of their depth and, near the end, less similar to early visual cortex and the other streams, under every metric we tried and unlike untrained decoders of the same depth. The other features of the raw trajectories—the size of the late dip in the ventral and lateral streams, a rebound at the returned state, and several checkpoint-specific exceptions—depend on feature scaling.

The returned final state needs care for two separate reasons. Architecture explains why it is unread: in both the causal and the prefix-LM decoders of our panel no block follows the last one, and next-token prediction reads only the final text position, so the last block's outputs at image positions feed no later computation. Architecture does not explain why alignment changes there. The output normalization rescales the few high-magnitude channels that dominate the last read state, and standardizing features removes most of the terminal change. Layer-wise comparisons that include the returned state should therefore report the last read state and a scale-invariant metric beside it. More generally, resemblance to cortex does not show that a state is used: model-side work finds that cross-modal reading concentrates in middle layers (Kaduri et al., 2025) and that late visual states

become interchangeable (Ghosh et al., 2026). Whether the late decline of early-cortex alignment occurs in states that answers depend on is a direct, testable next question.

Cortical RSA adds an anatomically differentiated reference to model-side evidence that visual tokens change their accessibility and role over depth (Neo et al., 2025; Liu et al., 2025; Fan et al., 2026), but it does not identify a mechanism: a falling early-cortex score may reflect selective loss of low-level relations, a basis change, or movement toward another geometry, and a rising lateral score does not show that a decoder implements lateral cortex. NOD changes participants, stimuli, field of view, preprocessing and cortical surface, so the 19/20 directional agreement (14/20 with standardized features, where the early decline is uncertain on NOD) is useful external evidence, while the wider NOD intervals show the cost of generalizing jointly over people and a smaller image set.

Our novelty is a measurement and empirical-synthesis contribution. Layer-wise VLM–brain comparisons (Bavaresco et al., 2026), layer- and region-resolved encoding of multimodal models on NSD (Oota et al., 2025), broad NSD model panels (Shah et al., 2026), and model-side visual-token trajectories (Venhoff et al., 2025; Krojer et al., 2026) already exist, and rogue dimensions are known to distort similarity in language models (Timkey & van Schijndel, 2021). We combine complete image-position trajectories with region-specific cortical RDMs, a fixed multi-lineage panel, an independent fMRI dataset, untrained-decoder controls, and a depth decomposition that separates read from unread states and metric effects from geometry. Oota et al. (2025) find that later layers of text-response embeddings align best with early visual regions in two of three models, whereas early-cortex alignment of image-token states falls late in our decoders; the readouts differ, and which token positions carry early visual geometry late in the stack is itself an open question.

Several limitations define the scope. First, 13 checkpoints are not an unbiased sample of VLMs; related checkpoints share decoders, vision towers, data, and training recipes. Lineage-balanced summaries reduce numerical dominance but do not create independent model replicates. Native image resolution, token count, prompt wrapper, attention mask and training data also vary across checkpoints, so apparent scale effects are not controlled interventions. Second, the participant samples are small, so signed-rank tests have coarse resolution and bootstrap intervals describe participant uncertainty rather than all sources of variation. Third, correlation-distance RSA discards response amplitude and can change under alternative similarity measures (Soni et al., 2024); we report four feature treatments, but whitened or cross-validated distances remain untested. Fourth, three large models lack NOD readouts, and older NOD caches encode their stimulus order through the extraction program rather than embedded filenames. Fifth, the depth decomposition, its 0.75-depth boundary and the metric controls were chosen after the curves had been inspected, the untrained controls cover nine of the thirteen checkpoints, and we did not measure how much each read state contributes to the model's answers. Finally, exact checkpoint revisions were not recorded during the original GPU extraction; the released cache hashes make the reported analysis auditable, but a fresh extraction from a moving repository is not guaranteed to be bit-identical.

The practical lesson is that brain alignment of a VLM cannot be summarized by its visual encoder or final layer alone: at image positions the final layer is not a state the model reads, and its alignment is set in large part by the output normalization. Reporting complete trajectories, the states a model reads, a scale-invariant metric, and counterexamples turns these choices from hidden researcher degrees of freedom into testable properties of multimodal representation learning.

AI USE STATEMENT

We used Anthropic Claude and OpenAI Codex to assist with code generation (including the data-preprocessing code) and debugging, methodological and statistical auditing, hypothesis refinement, results interpretation, literature-search triage, figure preparation, and manuscript drafting and editing. We did not use generative AI to create synthetic data, to prove mathematical claims, or to translate text, and no AI system generated, altered, or selected brain data or reported results: every reported result is written by scripts from the hashed result artifact. The authors reviewed the underlying sources, inspected all material code changes, reran the reported analyses and tests, and take responsibility for every claim and artifact in this submission.

ETHICS STATEMENT

This work reanalyses publicly available, de-identified human fMRI datasets under their respective data-use terms. Participant recruitment, informed consent, and ethical approval were handled by the original NSD and NOD studies. We redistribute no brain data and make no attempt to identify participants or infer sensitive individual attributes.

REPRODUCIBILITY STATEMENT

All analysis code, configuration files, figure-generating scripts, the compact hashed primary result with its saved bootstrap draws, and the generated values and tables are in the anonymous supplementary material and will be released publicly upon publication. The supplement specifies the frozen checkpoint panel, preprocessing and inference decisions, the software environment, and one command per figure and table; unit tests cover the RDM, preprocessing, provenance, and statistical utilities. The result artifact records cryptographic hashes of all numerical inputs and the package versions; its Git revision is withheld from the anonymous copy. The Natural Scenes Dataset is available at `http://naturalscenesdataset.org` under its data-use agreement and NOD as OpenNeuro dataset `ds004496`; we redistribute no brain data. Model weights are the public checkpoints named in Section 3; their exact repository revisions were not recorded during the original extraction, a limitation the artifact states explicitly.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report, 2025.

Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. Vision-language models align with human neural representations in concept processing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3255–3274, 2026. doi: 10.18653/v1/2026.eacl-long.150.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, 2025. doi: 10.1038/s42256-025-01072-0.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11987–11997, 2026.

Sambit Ghosh, R. Venkatesh Babu, and Chirag Agarwal. Do vision language models need to process image tokens?, 2026. Oral at the CVPR 2026 TRUE-V Workshop.

Zhengxin Gong, Ming Zhou, Yuxuan Dai, Yushan Wen, Youyi Liu, and Zonglei Zhen. A large-scale fMRI dataset for the visual processing of naturalistic scenes. *Scientific Data*, 10:559, 2023. doi: 10.1038/s41597-023-02471-x.

Larissa Höfling, Matthias Tangemann, Lotta Piefke, Susanne Keller, Katrin Franke, and Matthias Bethge. Only brains align with brains: Cross-region alignment patterns expose limits of normative models. In *International Conference on Learning Representations (ICLR)*, 2026.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? A deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. LatentLens: Revealing highly interpretable visual tokens in LLMs, 2026.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n. arXiv:2408.03326.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025. doi: 10.48550/arXiv.2510.04819.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv 2024; published ICLR 2025.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10(4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Subba Reddy Oota, Akshett Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi, Manish Shrivastava, Maneesh Singh, Bapi S. Raju, and Manish Gupta. Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain). In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2505.20029.

Arun Ram Ponnambalam and K. Pottore Venkiteswaran. Semantic information orthogonal to visual features peaks in lateral occipitotemporal cortex. *bioRxiv*, 2026. doi: 10.64898/2026.03.14.711805.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11: e77599, 2022. doi: 10.7554/eLife.77599.

Heiko H. Schütt, Alexander D. Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *eLife*, 12:e82566, 2023. doi: 10.7554/eLife.82566.

Arya Shah, Vaibhav Tripathi, Mayank Singh, and Chaklam Silpasuwanchai. Gaslight, gatekeep, V1-V3: Early visual cortex alignment shields vision-language models from sycophantic manipulation, 2026.

Ansh Soni, Sudhanshu Srivastava, Konrad P. Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024. doi: 10.1101/2024.08.07.607035. URL https://www.biorxiv.org/content/10.1101/2024.08.07.607035v1. Version 1; the unversioned DOI now resolves to a retitled 2026 revision.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, et al. PaliGemma 2: A family of versatile VLMs for transfer, 2024.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021. doi: 10.18653/v1/2021.emnlp-main.372.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal LLMs, 2025. arXiv:2506.11976v2.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A Audit trail and additional methodological detail

A.1 Brain-preprocessing correction

The corrected NSD cache is versioned `zscore_per_vertex_full_session_ddof0_v1`. Each of the 750 trial rows in a session contributes to its vertex-wise mean and population standard deviation; selected shared-image rows are written only after normalization. Statistics are accumulated in float64, zero-variance or non-finite vertices become non-finite cache columns, and the cross-participant mask removes those columns identically from every subject. Cache files carry the preprocessing version, source beta version, subject, session count, and row/ROI metadata. Publication is atomic through a temporary file so an interrupted transfer cannot masquerade as a complete subject.

The original NOD reliability diagnostic contained a transpose: it correlated grayordinates across images and returned one value per image rather than one value per grayordinate. The corrected implementation keeps arrays in image-by-grayordinate orientation, correlates odd- and even-run responses across the 120 images for every grayordinate, and then applies the Spearman–Brown correction. Unit tests fail if the output dimension is not the grayordinate dimension. Reliability is a reported QA measure; ROIs are not selected according to the observed model effect. The released NOD GLM betas are not standardized: their run-wise grayordinate scales vary substantially. The corrected cache therefore standardizes each grayordinate over the complete 120-image run before averaging the ten repetitions, and records this operation under a versioned normalization identifier.

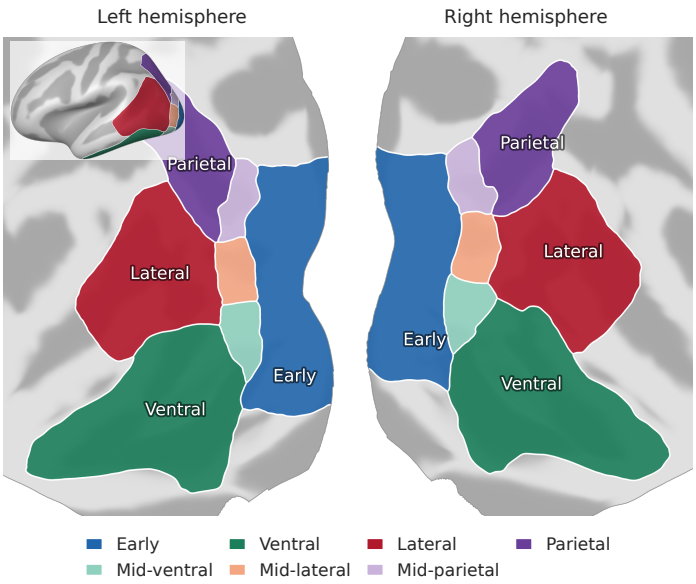

Figure 4: The seven `streams` regions of NSD on the flattened fsaverage surface of both hemispheres, cropped to posterior cortex, in the colours used in every figure. Early visual cortex lies in the middle, and the ventral, lateral and parietal streams extend from it. White gaps are the cuts that let the surface lie flat. Inset: the same regions on the inflated left hemisphere, seen from the side. All region-level alignment in this paper is measured within these parcels.

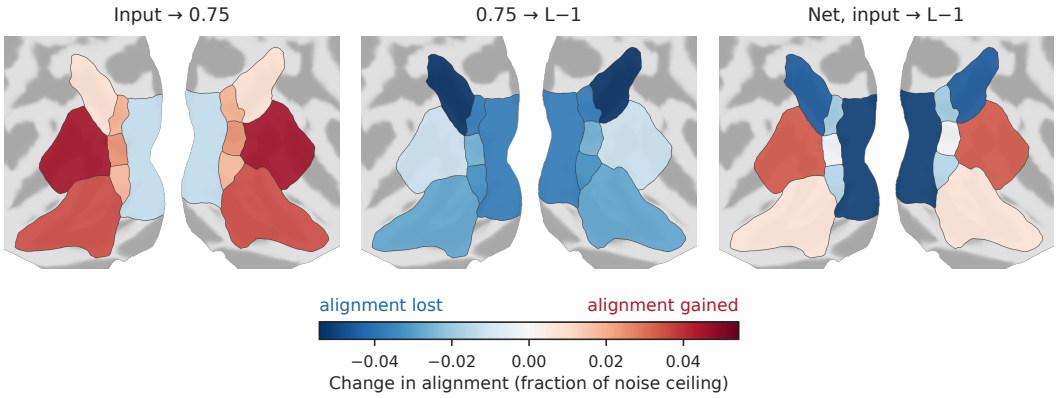

Figure 5: Depth steps of alignment on the flattened cortex. Each stream is coloured by its checkpoint-weighted change on NSD with standardized features: up to 0.75 depth (left), from there to the last read state (middle), and in total at the last read state (right).

A.2 THE CORTICAL REGIONS

A.3 METRIC CONTROLS ACROSS THE STREAMS

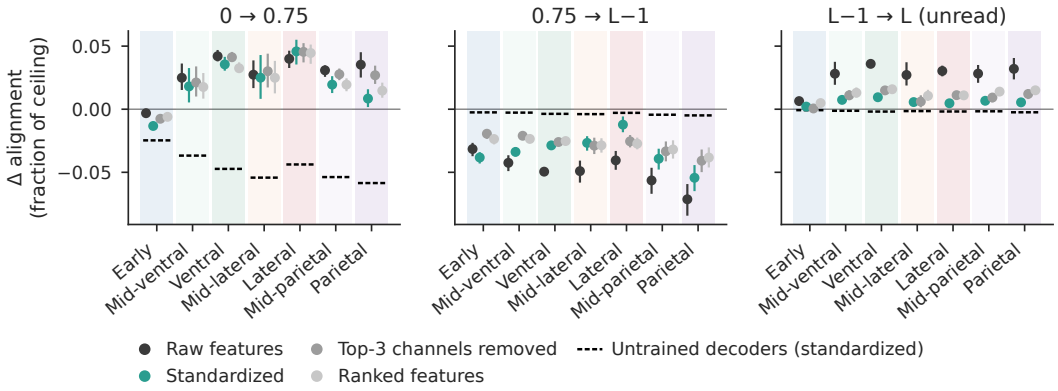

Figure 6: Depth steps of alignment for all seven streams on NSD under four feature treatments: raw (the frozen protocol), standardized across images, without the three highest-energy channels, and ranked. Dots: checkpoint-weighted participant means with 95% participant-bootstrap intervals; dashed lines: depth-matched untrained decoders (9 checkpoints, standardized features).

A.4 PER-CHECKPOINT DEPTH PROFILES

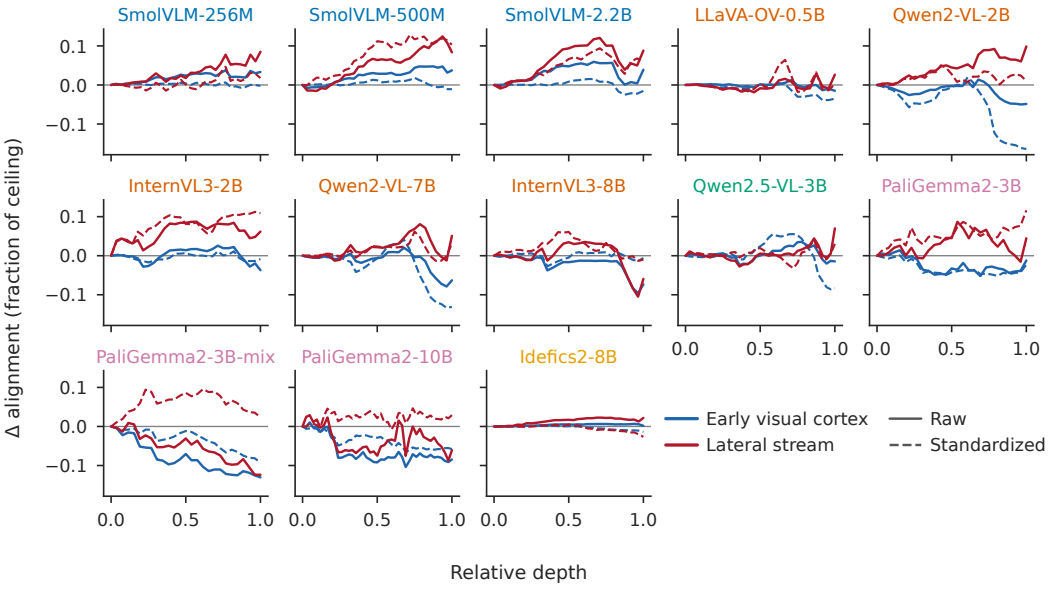

Figure 7: The full depth trajectory of every checkpoint, one panel each: change in early-cortex (blue) and lateral-stream (red) alignment from the decoder input, as a fraction of the lower noise ceiling, with raw (solid) and standardized (dashed) features. Panel titles are coloured by decoder lineage. This is the per-checkpoint expansion of Figure 2.

Table 1: Corrected endpoint change (last decoder readout minus `llm.00`), as a fraction of the lower RSA noise ceiling. Bold values have a 95% subject-bootstrap interval excluding zero. Checkpoints are a fixed panel; dashes mark missing NOD representations.

	NSD		NOD	
Checkpoint	Early	Lateral	Early	Lateral
SmolVLM-256M	**+0.033**	**+0.085**	**+0.059**	**+0.200**
SmolVLM-500M	**+0.037**	**+0.084**	+0.018	**+0.196**
SmolVLM-2.2B	**+0.039**	**+0.088**	+0.024	**+0.173**
Qwen2-VL-2B	**-0.049**	**+0.099**	**-0.111**	**+0.178**
LLaVA-OV-0.5B	**-0.015**	**+0.026**	+0.006	**+0.052**
InternVL3-2B	**-0.037**	**+0.062**	**-0.059**	**+0.070**
Qwen2.5-VL-3B	**-0.015**	**+0.070**	**-0.051**	**+0.082**
PaliGemma2-3B	**-0.012**	**+0.045**	**-0.035**	**+0.061**
PaliGemma2-3B-mix	**-0.130**	**-0.124**	**-0.165**	**-0.162**
PaliGemma2-10B	**-0.085**	**-0.061**	—	—
Idefics2-8B	+0.002	**+0.022**	—	—
Qwen2-VL-7B	**-0.063**	**+0.051**	—	—
InternVL3-8B	**-0.074**	**-0.059**	**-0.080**	**-0.080**

A.5 Additional endpoint results

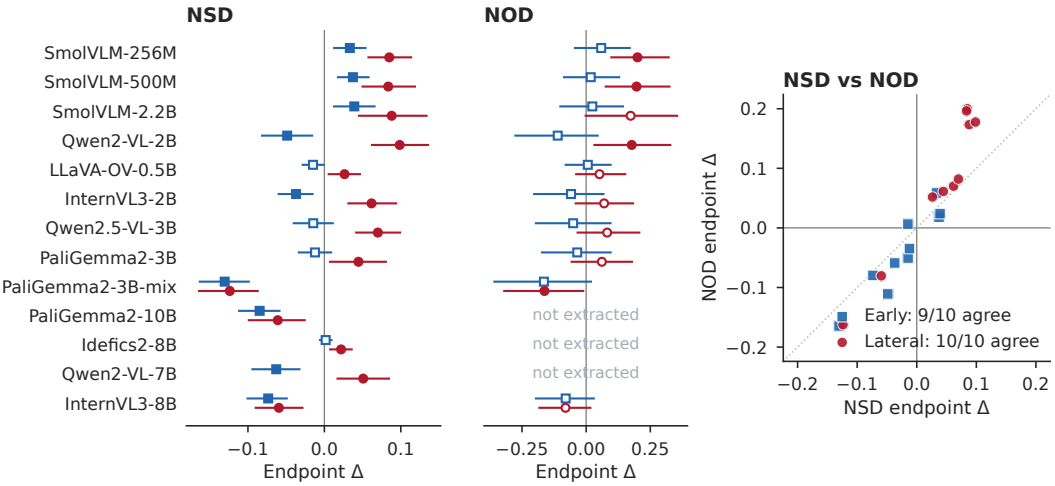

Figure 8: Endpoint change of every checkpoint on NSD and NOD (squares: early cortex; circles: lateral stream) with 95% participant×stimulus intervals; filled markers exclude zero. Right: NSD against NOD for the checkpoints available on both. These are fixed-panel descriptions, not draws from a population of VLMs.

A.6 The InternVL3 scale contrast

Before the NSD preprocessing correction, InternVL3-8B appeared to reverse both regional directions relative to InternVL3-2B. That preliminary result does not survive the corrected endpoint analysis: early alignment decreases at both scales. The remaining scale contrast is lateral, positive for 2B and negative for 8B on NSD. Six extraction invariants on the 8B checkpoint rule out a shifted image-token mask, an incorrect number of hidden states, and non-image leakage at the embedding readout; mean, trimmed-mean and median pooling retain the 8B lateral sign, and reduced-precision RDMs closely match float32. Removing the final 10% or 25% of states makes the 8B lateral trend positive on both NSD and NOD, so the reversal is a terminal-state property of this checkpoint on both image sets. One 2B–8B contrast is not evidence that scaling InternVL generally reverses cortical trajectories.

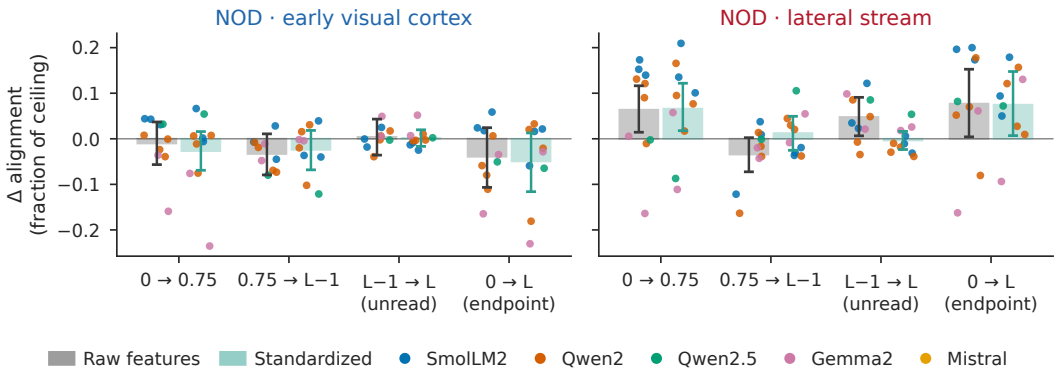

Figure 9: The endpoint decomposed by depth on NOD, as in Figure 3.

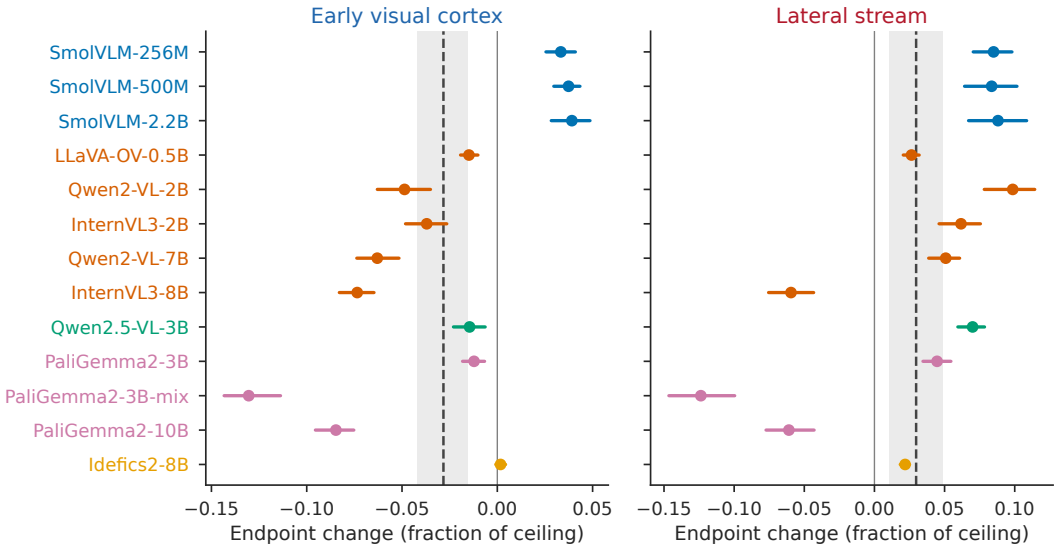

Figure 10: Per-checkpoint endpoint change with 95% subject-bootstrap intervals (dot: mean over the eight participants; bar: interval), grouped by decoder lineage. The dashed line and grey band are the checkpoint-weighted panel estimate and its 95% subject×stimulus interval. **Left:** early cortex. **Right:** lateral stream.

A.7 Analysis versions

The frozen specification of version 1 (3 September 2026) is kept unchanged together with its input manifest. Version 2 (14 September 2026) followed an independent audit and changed three things without touching the estimand, panel, pooling, regions, preprocessing or tests. Spearman noise ceilings pool ranked rather than raw RDMs for NSD and NOD, which moves ceiling-scaled values by at most about 5%. The joint participant–stimulus bootstrap uses 10,000 saved draws instead of 1,000, because the version-1 lower bound of the NOD lineage-balanced lateral interval was not stable at 1,000 draws. Every participant's curve is stored. Version 2 also added the exploratory depth decomposition of Section 4.3; its estimands draw from a separate random generator, so adding them alters no other interval. The standardized rerun of Section 4.4 is the version-2 code with one switch, standardizing every model feature across images before its RDM is formed, and is reported as exploratory.

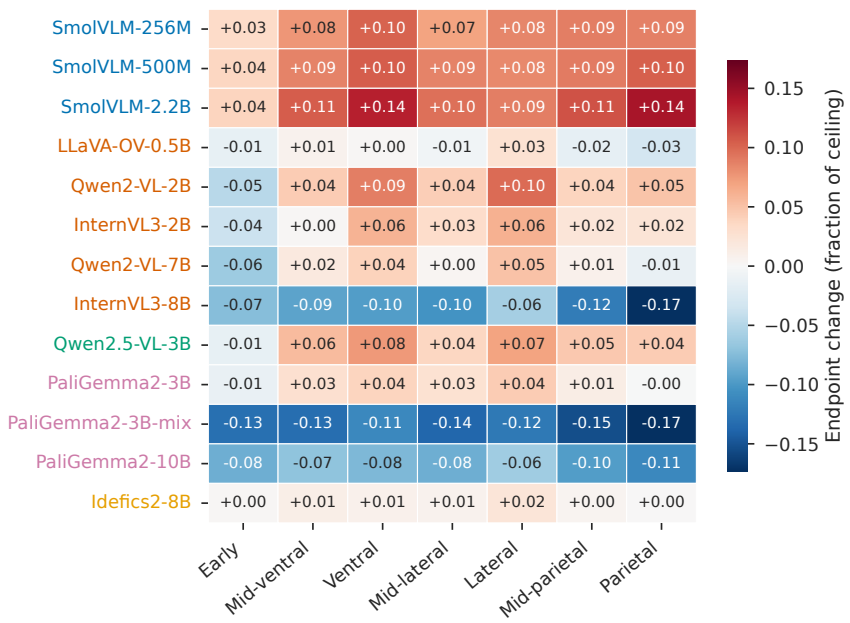

Figure 11: The endpoint contrast for every checkpoint against all seven `streams` regions (blue: alignment declines through the stack; red: rises). Row labels are coloured by decoder lineage; the numeric values are in Table 2.

Table 2: Endpoint alignment change (last language-model readout minus `llm.00`) for every checkpoint against all measured cortical regions, as a fraction of the lower RSA noise ceiling (NSD, mean pooling, full-session z-scored betas). Bold marks a 95% subject-bootstrap interval excluding zero. Checkpoints are a fixed panel grouped by decoder lineage; the effect is regional rather than global and heterogeneous across checkpoints.

Checkpoint	Early	MidV	Vent	MidL	Lat	MidP	Par
SmolVLM-2.2B	**+0.039**	**+0.106**	**+0.135**	**+0.096**	**+0.088**	**+0.111**	**+0.142**
SmolVLM-256M	**+0.033**	**+0.076**	**+0.102**	**+0.070**	**+0.085**	**+0.088**	**+0.087**
SmolVLM-500M	**+0.037**	**+0.085**	**+0.105**	**+0.086**	**+0.084**	**+0.089**	**+0.103**
LLaVA-OV-0.5B	**-0.015**	+0.008	+0.002	-0.006	**+0.026**	**-0.016**	**-0.032**
Qwen2-VL-2B	**-0.049**	**+0.044**	**+0.089**	**+0.043**	**+0.099**	**+0.039**	**+0.050**
InternVL3-2B	**-0.037**	+0.002	**+0.060**	**+0.032**	**+0.062**	**+0.022**	**+0.024**
Qwen2-VL-7B	**-0.063**	+0.017	**+0.040**	+0.004	**+0.051**	+0.008	-0.013
InternVL3-8B	**-0.074**	**-0.093**	**-0.096**	**-0.104**	**-0.059**	**-0.121**	**-0.173**
Qwen2.5-VL-3B	**-0.015**	**+0.055**	**+0.076**	**+0.038**	**+0.070**	**+0.047**	**+0.044**
PaliGemma2-3B	**-0.012**	**+0.027**	**+0.039**	**+0.026**	**+0.045**	+0.012	-0.001
PaliGemma2-3B-mix	**-0.130**	**-0.131**	**-0.112**	**-0.137**	**-0.124**	**-0.154**	**-0.173**
PaliGemma2-10B	**-0.085**	**-0.070**	**-0.078**	**-0.084**	**-0.061**	**-0.097**	**-0.112**
Idefics2-8B	+0.002	**+0.012**	**+0.009**	**+0.008**	**+0.022**	**+0.004**	+0.001

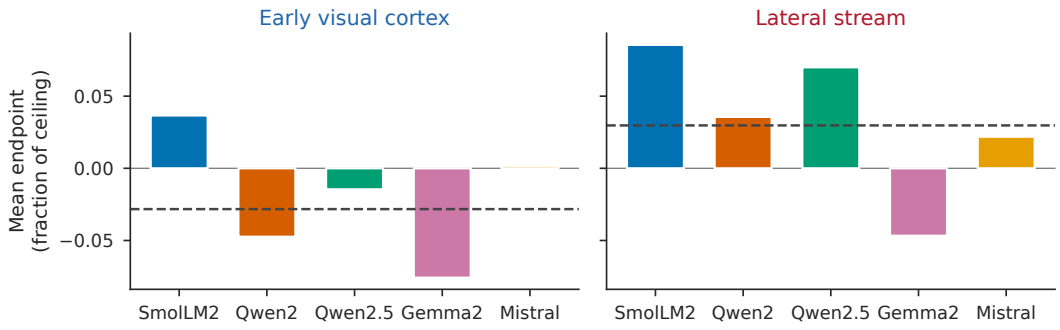

Figure 12: Effect by decoder lineage, each lineage summarised as the mean endpoint over its checkpoints (early cortex left, lateral stream right). The dashed line is the fixed-panel mean.

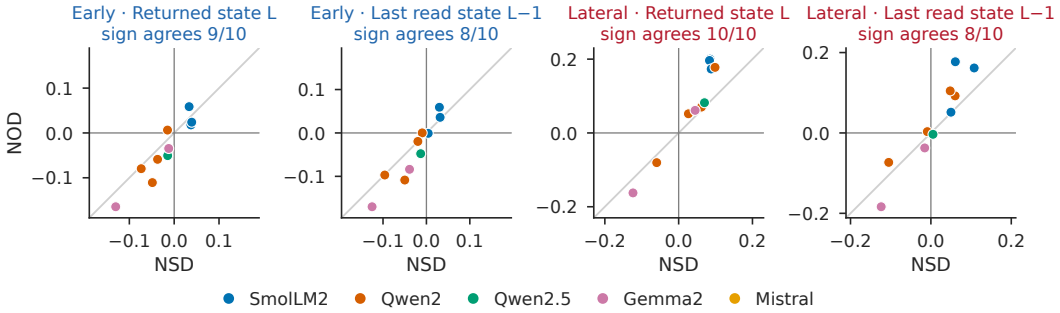

Figure 13: Cross-dataset comparison. Each point is one checkpoint's change on NSD (horizontal) against NOD (vertical), for early cortex and the lateral stream, measured at the returned state L and at the last read state $L-1$; points on the same side of both zero lines agree in direction. Colour is decoder lineage.

A.8 ATLAS-TRANSFER QA

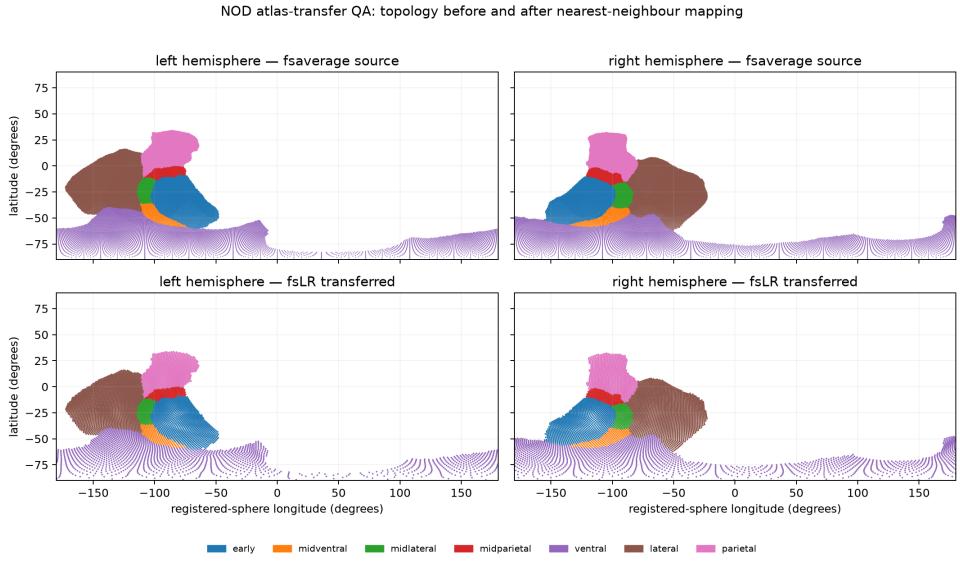

Figure 14: Quality assurance for transferring the seven NSD stream labels from fsaverage to the NOD fsLR32k surface. The transfer uses the registered spherical coordinates and nearest-neighbour assignment separately by hemisphere. Label coverage and regional shares are checked against the source atlas before any model comparison.

A.9 RESULT ARTIFACT

The machine-readable primary result file contains, for each dataset, checkpoint, pooling rule, ROI, and scaling: all participant trajectories; endpoint, relative-depth slope, and depth-rank summaries; 10% and 25% terminal exclusions; bootstrap intervals; sign counts; signed-rank tests; and the relevant Holm-adjusted values. It also contains the checkpoint- and lineage-weighted panel summaries, leave-one-lineage-out descriptions, the depth-step estimands with their panel summaries, ROI ceilings and NOD reliability values. A manifest records SHA-256, byte size, and relative path for the analysis script and its imported local modules, environment lockfile, frozen specification, model configurations, NSD design, corrected subject caches, NOD brain and atlas files, and every representation cache used. Tables and the primary figure are generated from this file and refuse an unhashed input.

The anonymous supplement includes this compact result JSON with its saved bootstrap draws, the standardized rerun with its draws, the metric and untrained-decoder controls, the depth profiles behind every curve, generated LaTeX values and tables, the vector figures, and atlas QA together with source, tests, configurations, lockfile, and a second manifest. Restricted fMRI data, stimuli, model weights, and multi-gigabyte activation caches are excluded.

A.10 CONTROLS OUTSIDE THE CONFIRMATORY ESTIMAND

The repository also implements: (i) mean/trimmed/median image-token readouts; (ii) a generation-position readout; (iii) two reduced-precision formats against float32; (iv) depth-matched randomly initialized decoder controls; (v) image-token attention mass with a causal-prefix baseline; (vi) image-token mean-replacement interventions; (vii) low-level RDM partialling; and (viii) a participant-by-stimulus bootstrap. These analyses were developed at different stages of exploration. They are useful diagnostics but are not promoted to co-primary hypotheses after their outcomes were observed. The frozen endpoint matrix, its declared sensitivities, and the independent NOD comparison determine the headline conclusions.

A.11 KNOWN REPRODUCIBILITY BOUNDARIES

All numerical inputs to the CPU analysis are content-addressed. However, most VLM caches were extracted before immutable Hugging Face weight and processor revisions were recorded; the result registry therefore says `not_recorded_at_representation_extraction` rather than inventing revisions. Ten NOD checkpoints are analyzed, but the three missing large-model caches cannot be reconstructed on the local 16-GB GPU in the same float32 protocol. In addition, all ten analyzed NOD checkpoint caches predate embedded image-name metadata. Their order follows the brain-cache name list in the extraction program, but the primary result labels this weaker provenance explicitly. These are limitations of fresh extraction, not hidden substitutions in the reported computation.