DOES THE WORKSPACE SEE? TESTING A MULTIMODAL GLOBAL WORKSPACE IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Jacobian lens identifies a small, workspace-like subset of internal representations through their average influence on model outputs. A readout indexed by the output vocabulary can only return tokens, so for vision-language models we cannot tell whether visual content lies inside the subset or outside it. We call the question of whether a representation can be expressed in the symbols of a given output channel *channel expressibility*. Verbalizability, the property the lens reports for the subset, is the special case where that channel is text; a single-channel model cannot test it, since its readout has only that alphabet. We introduce MJ-lens, a modality-split Jacobian lens (four operators in a unified model, two in an understanding-only model) that also adds a readout into the discrete visual codebook by linearizing the image-generation head. We test whether the workspace account extends to multimodal models, using understanding-only and unified models. We find that the subspace MJ-lens selects has a causal role and its own transport geometry, but it does not satisfy all workspace signatures. In the understanding-only model, ablating this subspace impairs some visual judgments, but nameability does not predict which ones. In the unified model, the output channel changes what the subspace contains, yet overwriting a lens coordinate does not change what the model draws. Directions obtained without the lens do change it, installing a specified object only at early layers and a colour across a broad middle band, while the lens's own coordinates move nothing. The bound is on the instrument, not the representation: the lens can read the subspace, but as a readout operator it does not provide directions that control what the model draws.

1 INTRODUCTION

Interpretability has a recurring problem: when internal states are read through the output vocabulary, the readout always returns vocabulary tokens, even when the states themselves are not encoded in that form. The Jacobian lens (J-lens; Gurnee et al., 2026) is the strongest recent example of this approach and is unusually well validated for text models. It defines the meaning of an activation as its averaged first-order effect on the final output distribution, recovers a small selected subset of directions, and shows that this subset behaves like a global workspace in the functional sense: its contents can be reported, held, and operated on by whatever computation the context supplies, while much automatic processing proceeds without them. The work also reports that the subset is verbalizable.

Verbalizability differs from these functional criteria. In a single-channel text model, the output vocabulary is the only available alphabet. The readout is indexed by it, so every direction the lens recovers can be expressed in words. This is a consequence of the definition, not empirical evidence. We call this *single-channel definitional closure* (§2.2). It leaves undecided whether workspace content is expressed in words because the workspace has a verbal format, or because the lens can only read through the output vocabulary. This question is a special case of a more general one: whether a representation can be expressed in the symbols of a given output channel. We call this property **channel expressibility**. Deciding it requires a second output channel, for example the image-generation channel of a unified model. The original J-space paper (Gurnee et al., 2026) does not examine the

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.

multimodal case in depth, and no published work applies the Jacobian lens to a vision-language model.

Closure also means that asking whether some visual content lies inside the subset returns the definition rather than a measurement. We replace that question with three that require empirical test and admit negative answers. (i) What structure do the transport operators and activation statistics show at visual positions? (ii) Which visual tasks fail when the subspace is removed? (iii) When the model is about to output an image rather than text, what does the subset contain, and what does it control under intervention? The third question requires a unified model. Its image codebook provides the second alphabet this paper needs. We linearize its image-generation head into an *effective visual unembedding*, so that the same activation can be read out over both alphabets and rendered through the model's own decoder as a picture (§2.3).

We do not test whether any content belongs to the workspace. We test channel expressibility, causal dependence, and interventional writability. The claims that result are of three kinds, held to different standards. Geometric and statistical claims (§3.2) do not depend on what the readout's tokens mean. Causal-ablation claims (§3.3) depend on behaviour. Readout-semantic claims are bounded by the **faithfulness band**, the depth range in which the readout has a measurable correspondence to the model's final output (§3.1).

Contributions. We test whether the workspace account extends to multimodal models. We find that the subspace selected by MJ-lens has a causal role and its own transport geometry, but it does not satisfy all workspace signatures. The central claim of the paper concerns the lens: **the Jacobian lens, as a readout operator, does not provide control directions.** Visual content at middle depths can be read in the image channel and written by directions obtained without the lens, its appearance at middle depths and an object's identity only at early layers, but the coordinates the lens selects cannot write it. This is a limitation of the reading lens, not necessarily a read–write gap of the workspace itself.

(1) On Janus-Pro (Chen et al., 2025), J-lens directions do not control image generation, but some non-lens directions do (§3.4–§3.5). Directions obtained without the lens, for example difference-of-means directions (Panickssery et al., 2023), can be injected into the residual stream to causally change the object and colour the model draws. The lens's own selected coordinates cannot. This positive control places the failure in the instrument rather than in the visual representation. **(2) MJ-lens makes channel expressibility measurable** (§2.3). A modality-split Jacobian lens, together with the faithfulness band, turns channel expressibility from a property fixed by construction into a measurable one. **(3) Nameability does not determine which visual tasks depend on the subspace** (§3.2, §3.3, §3.4). The workspace account predicts that only nameable content can enter the workspace and depend on it for processing. But the experiments find that which visual judgments depend on the lens-selected subspace is not predicted by whether the judged property is nameable. In the unified model, the content of this subspace also varies with the modality the model is about to output.

Related work. Gurnee et al. (2026) introduce the Jacobian lens and the J-space in text-only models; we extend the lens to multimodal models. Vocabulary readouts (nostalgebraist, 2020; Belrose et al., 2023) are known to be unreliable in early layers (Yom Din et al., 2024; Phukan et al., 2025); our faithfulness band measures that range for our readout. Work on visual tokens in VLMs characterizes their representations (Neo et al., 2025; Krojer et al., 2026), and Shahgir et al. (2026) relate visual failures to nameability through behaviour, which we test by ablation against a rotated control (Makelov et al., 2024). Because of the page limit, the full discussion is in Appendix A.

2 Problem Statement and Methodology

2.1 The Jacobian Lens and the J-space

The Jacobian lens (Gurnee et al., 2026) reads an activation $h_{\ell,t} \in \mathbb{R}^d$ in the residual stream through its averaged first-order effect on the final layer:

$$J_\ell \;=\; \mathbb{E}_{p,\,t,\,t'\geq t}\left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}}\right], \qquad \text{lens}_\ell(h) \;=\; \text{softmax}\big(W_U\,\text{norm}(J_\ell\,h)\big), \tag{1}$$

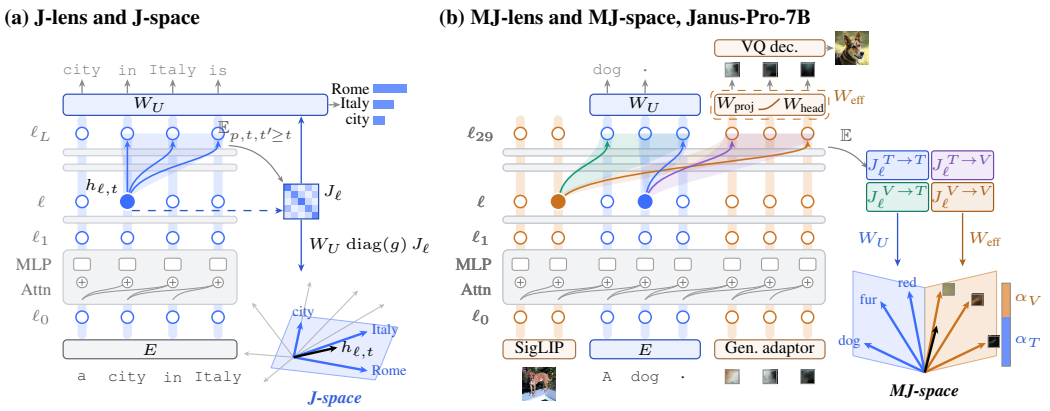

Figure 1: **The J-lens and J-space on a text-only transformer (a), and MJ-lens and the MJ-space on Janus-Pro-7B (b).** *(a)* Jacobians from $h_{\ell,t}$ to all later final-layer states are averaged into J_ℓ and decoded with W_U; the J-space is a small subset of the resulting vocabulary-indexed directions, verbalizable by construction. *(b)* MJ-lens averages the Jacobians separately for each pair of source and target modality, decodes text targets with W_U and image targets with the linearized image head W_{eff}, and decomposes an activation over both dictionaries (the MJ-space); the visual share of that decomposition is the modality index. Images are real: a COCO (Lin et al., 2014) input and tokens from a Janus-Pro-7B generation.

with the expectation over prompts p, source positions t, and current and future target positions t'. The transport J_ℓ is one $d \times d$ matrix per layer. The decoding reuses the model's own unembedding W_U. The readout therefore names the output tokens this activation would push the model toward. Its dictionary is the rows of $W_U \operatorname{diag}(g)\, J_\ell$, indexed by vocabulary entries. The RMSNorm (Zhang & Sennrich, 2019) gain g is load-bearing (Appendix D).

Applying Eq. (1) across layers, Gurnee et al. (2026) report that a small subset of directions, the J-space, carries most of the dispositional content in a middle band of layers and behaves like a global workspace. We use this term in the functional sense the original does. They characterize it by five functional criteria (verbal report, directed modulation, internal reasoning, flexible generalization, selectivity). They also give structural signatures (excess kurtosis of the coefficients, effective dimensionality, occupancy) that mark the band. The signatures transfer directly. The functional criteria do not, and §4 states which ones we test.

2.2 Channel expressibility

The rows of W_U are the vocabulary, so a readout built on W_U can only name tokens. A text-only model has no other output. Finding word-like content in its J-space therefore could not have failed. The property the original work (Gurnee et al., 2026) calls *verbalizability* is thus a special case of a more general one, which we call **channel expressibility**: whether a representation can be expressed in the symbols of a given output channel. A model with two output channels can be asked about each separately, so there the property can fail. Every experiment on the unified model below tests it in the image channel. Appendix B works the argument through.

A representation is expressible in a channel when it can be decomposed into that channel's symbols. For the text channel, the symbols are vocabulary entries and the decomposition is the lens readout itself. For the image channel, the symbols are codebook entries and the measure is the **modality index**: the fraction of reconstruction mass that lands on visual atoms, in excess of the same fraction under a rotated visual dictionary. Section 3.4 measures this quantity, and Section 3.5 tests whether the same content can be written back through the channel.

2.3 MJ-lens: modality-split transport and a second alphabet

Eq. (1) averages over all source and target positions. In a text model this is harmless, because every position carries the same kind of content. In a multimodal model, each position has a modality

$\mu(t) \in \{V, T\}$, and the pooled operator mixes four transports. We separate them:

$$J_\ell^{m \to m'} \;=\; \mathbb{E}_{p \sim \mathcal{D}_{\text{mm}}}\, \mathbb{E}_{(t,t') \,:\, \mu(t)=m,\ \mu(t')=m',\ t' \geq t} \left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}} \right], \qquad m, m' \in \{V, T\}, \tag{2}$$

estimated on an interleaved image–text corpus. An understanding-only model admits $J^{T \to T}$ and $J^{V \to T}$; a unified model additionally admits $J^{T \to V}$ and $J^{V \to V}$. §3.2 tests whether the split helps.

A unified model with a discrete visual codebook needs an image-side analogue of W_U. Janus-style models generate images through a head with a pointwise nonlinearity, which we linearize around its operating point:

$$W_{\text{eff}} \;=\; W_{\text{head}}\, \text{diag}\big(\mathbb{E}_t[\sigma'(z_t)]\big)\, W_{\text{proj}}, \qquad \text{lens}_\ell^{\text{vis}}(h) \;=\; \text{softmax}\big(W_{\text{eff}}\, \text{norm}(J_\ell h)\big), \tag{3}$$

which has the same form as the text readout. We verify W_{eff} before use rather than assume it (§3.4).

The visual readout is applied position by position. At each visual position we take the top-ranked codebook entry. Placing the 576 entries back in their positions and passing them through the VQ decoder (van den Oord et al., 2017) produces an image. Spatial structure therefore comes from the activations, not from the operator. The operator has no spatial index of its own. In particular, $J_\ell^{T \to V}$ averages over all target positions, so a readout from a single text position is spatially uniform. A structured image needs many source positions at once, or an operator resolved by target position (Figure 4c). At visual source positions the rendered display does exactly this. That is why it turns out to be an echo of the input rather than a visual description the model builds (§3.4).

Given both dictionaries, we take a sparse non-negative decomposition of an activation over $[V^{\text{text}}; V^{\text{vis}}]$. The share of reconstruction mass carried by visual atoms is the modality index of §2.2 (Figure 1b). The index has three controls: a rotated visual dictionary, a size-matched text dictionary and a split-half estimate. They are given in Appendix C.1.

We call the union of the two dictionaries the **MJ-space**. Its text part is spanned by the lens vectors $W_U\, \text{diag}(g)\, J^{m \to T}$ and its visual part by $W_{\text{eff}}\, \text{diag}(g)\, J^{m \to V}$, where $m \in \{T, V\}$ is the modality of the source position. An understanding-only model has no $J^{m \to V}$, so its MJ-space has no visual part. Whether the two parts are independent, in the sense that writing to one leaves the other's output unchanged, is tested for the image output in §3.5.

3 Experiments

3.1 Models, data, and measurement

We ask three questions about the subspace the MJ-lens selects. First, how does its transport geometry at visual positions compare with that at text positions? Second, which visual judgments depend on it? Third, when the model has a second output channel, what can be read out of it and written into it?

Results come from two model families. Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) is the understanding-only model, used in Sections 3.2 and 3.3. Qwen3-VL-32B (Bai et al., 2025a) checks the conclusions at a second scale. Janus-Pro (1B and 7B) is the unified model, used in Sections 3.4 to 3.5. Its image codebook provides the second alphabet. All operators are fitted with the estimator of Appendix D. Figure 2 shows one example of each kind of experiment.

Readout validity. A readout that looks plausible is not thereby trustworthy. Before using it, we measure the depth range over which the lens readout tracks the model's own output. We call this range the **faithfulness band**. It is set by the first of two checks. The range check compares the readout against the model's final-layer logits. The content check asks whether a particular reading is what the model computes with. Our readout passes the range check over L18–L27 on Qwen2.5-VL-7B-Instruct. It does not pass the content check. Appendix E (Figure 6) gives both checks and audits which claims in this paper fall outside the band.

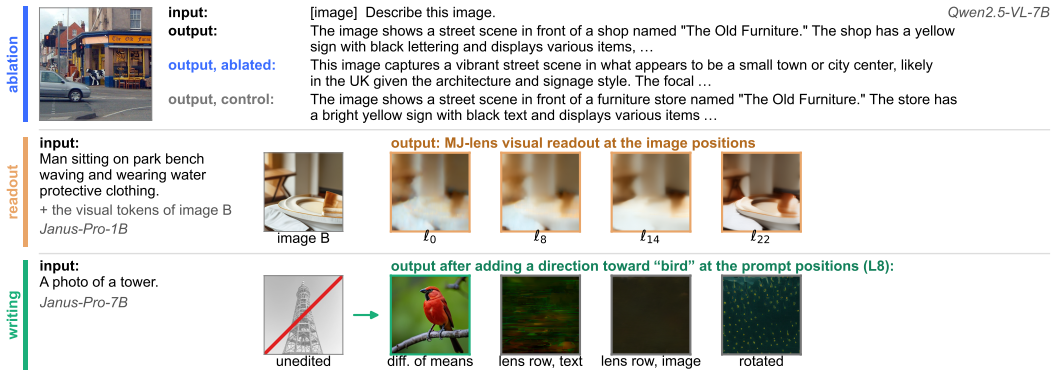

Figure 2: **One example of each experiment: ablation (§3.3), the rendered readout (§3.4), and writing into the image channel (§3.5).** *Top:* Qwen2.5-VL-7B, 25 lens-selected or random directions removed at visual positions (L9–L17): the ablated caption drops the furniture shop, the control keeps it. *Middle:* Janus-Pro-1B, the prompt of image A with the visual tokens of image B: the readout renders B at every layer. *Bottom:* Janus-Pro-7B, a direction toward "bird" added at L8: the difference of means draws a bird, the lens rows and the rotated control do not.

3.2 The visual side has its own transport geometry

This section measures the geometry of the lens operators. The visual transport operator differs from the text one. The lens-selected subspace at visual positions is built mid-network and consumed before the output. It is also broadcast along the token axis.

The modality split is necessary. We compare the visual-to-text and text-to-text operators with centered kernel alignment (CKA; Kornblith et al., 2019). They differ by more than sampling noise, and by more than twelve layers of depth do within one modality. Pooling the two modalities into one operator would therefore average over a difference larger than a large depth difference (Appendix F.3). Qwen3-VL-32B's $J^{V \to T}$ is under-calibrated through the first two fifths of the network, so its operator comparison is reported as undetermined.

The subspace is built mid-network and spent before the output. At visual positions the lens dictionary spans many directions mid-network and fewer near the output. At text positions it keeps expanding with depth (Figure 8a). The subspace also tracks a perceptual change as a step rather than in proportion. When we interpolate between two category images, its response jumps in the middle of the morph from L10 onward. The full residual stream changes smoothly at every depth (Appendix F.5).

Broadcast transfers along the token axis. Attention heads amplify lens directions more than a rotated control. The amplification is the same for visual and text directions (Appendix F.4). The subspace is therefore distributed to other positions, not only to the output. The other two broadcast signatures we tested do not survive their controls.

Visual positions therefore have their own transport geometry, distinguishable from the text operator by more than depth. This does not show that the subspace is a workspace.

3.3 Which visual tasks are impaired by ablating the lens-selected subspace is not predicted by whether their features are nameable

The original J-space paper (Gurnee et al., 2026) defines a workspace by five functional criteria. The most important is verbal report: when the model is asked what it is thinking about, it names concepts represented in the workspace. The MJ-lens gives us a subspace in a multimodal model, and we ask whether the same criterion holds there. We operationalize verbalizability in its sharpest case, a single word. A task is a two-alternative choice the model makes about an image, such as red against blue or left against right. We call the task *nameable* when the property that distinguishes the two alternatives, such as colour or orientation, can be named by a single word. The *verbalizability*

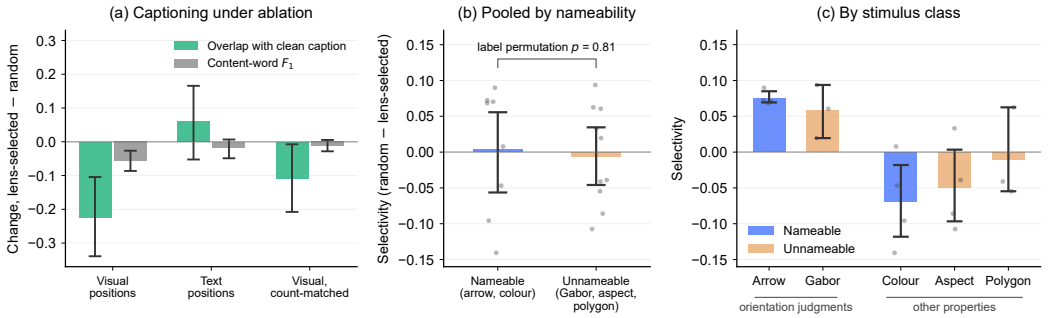

Figure 3: **Ablating the lens-selected subspace has a causal effect; the nameability contrast is null, and the extra loss sorts by task class.** *(a)* Captioning: paired change against removing as many random directions, in caption overlap (circles) and content-word F_1 (squares). *(b)* Selectivity, $\mathrm{AUC}_{\text{random}} - \mathrm{AUC}_{\text{lens-selected}}$, pooled by nameability. *(c)* The same by stimulus class, grouped by the property judged. Filled markers are nameable, open unnameable, grey dots difficulty levels; intervals are bootstrap 95% CIs.

hypothesis is that tasks supported by a verbalizable property are impaired when the subspace is ablated. We test it in three steps. Does ablation impair behavior at all? Does nameability determine which tasks are impaired? What does the extra loss actually sort by?

The subspace has a causal role. We let the model caption an image. We ablate the lens-selected subspace (the directions that sparse non-negative pursuit selects from the lens dictionary) at positions that carry visual tokens, and measure how much the caption changes. The control removes an equally large set of random directions from the same dictionary. Ablation changes the caption more than the control does. Overlap with the clean caption falls by -0.224 $[-0.339, -0.105]$, and content-word F_1 against the human captions falls by -0.057 $[-0.086, -0.026]$. The same ablation at positions that carry text tokens has no such effect ($+0.059$ $[-0.053, +0.166]$). The effect survives a count-matched control, though more weakly (Figure 3a; setup in Appendix G.1).

Nameability does not determine which tasks are impaired. We judge a task nameable by three conditions. First, its discriminative property has a name whose center word is a single token, so the name can be one vocabulary entry. Second, independent annotators give the same name, so the name is conventional. Third, a text-only model given only the name and shared context can solve the discrimination, so the name carries enough information for the task (Appendix C.2; difficulty calibration in Appendix G.2).

The comparison of nameable and unnameable tasks is null. Nameable tasks average $+0.003$ and unnameable tasks -0.006, a difference of $+0.009$ against a permutation null with standard deviation 0.036, $p = 0.81$ (Figure 3b). Teaching the model an arbitrary name for an unnameable feature does not raise its ablation sensitivity above an irrelevant-name control either ($+0.015$ against $+0.025$, Appendix G.3). Nor does reasoning depth decide it. On nine visual tasks, those that require relational or multi-step reasoning are not harmed more than those that need a single low-level judgment ($p = 0.97$, Appendix G.2). Task lists and prompts are in Appendices G.2 and G.2.

The extra loss sorts by task class instead. When we average ablation damage over the two groups, they look the same. When we look at each task class separately, they do not (Figure 3c). Arrow orientation is harmed at every difficulty level, and so is Gabor orientation, one of our unnameable classes. Colour runs the other way. In the nine-task battery, ablating the lens-selected directions at visual positions drives a difficulty-calibrated brightness comparison from AUC 0.899 to chance, 0.514, while the control stays at 0.889 (Figure 11). The extra loss is $+0.375$ $[+0.19, +0.59]$ (Appendix G.2). The classes with an extra loss are both orientation judgments. Those without are colour, aspect ratio and polygon shape. We therefore record a hypothesis for later work, not a finding. Whether ablation hurts a task may depend on which visual property it asks about, rather than on whether that property can be named. With five classes the grouping could be chance, and no experiment here was designed to test it. A cheap decisive test is to cross perceptual dimension with question form; we have not run it.



Figure 4: **At visual token positions the rendered readout echoes its input.** Janus-Pro-1B; every panel is decoded by the model's own VQ decoder. *(a)* Rendered readout at visual token positions: the scene resolves with depth toward the emitted image. *(b)* Swap control, prompt of caption A with the visual tokens of image B: every layer, including L0, renders image B. *(c)* At a text position, with the transport resolved by target position (L16): a dog image pushed along the transported "apple" direction (top) or a random direction of the same norm (bottom), by α times the local activation norm; α=0 is the model's own image. Both survive small pushes and break at large ones; neither turns into an apple.

What this branch establishes, and what it does not. Our experiments do not support the verbalizability hypothesis. They do not show that the MJ-lens is unreliable. The subspace it selects has a clear causal effect. They show that the first functional criterion of the original work does not extend to the multimodal setting in the form we tested (discussion in Appendix E).

3.4 What the image channel reads out

Sections 3.2 and 3.3 used an understanding-only model, which has no image output. We now turn to a unified model, Janus-Pro, whose image codebook gives us a second output channel. This section tests channel expressibility in the image channel. We ask whether a representation can be expressed in the image codebook, and how much of it can be. We report two results. At positions that carry visual tokens, the rendered readout is an echo of its own input. At text token positions, the visual component of the subspace is large when the model is about to emit an image, and small when it is about to emit text. By *rendered readout* we mean the top-ranked codebook entries passed through the model's own VQ decoder, so that the readout is shown as an image rather than a word (Appendix H.1).

At visual token positions, the rendered readout is an echo of its own input. The rendered scene starts blurred and resolves with depth (Figure 4a), as it would if the model were accumulating a visual description. A swap shows otherwise. We pair the visual tokens of image B with a caption that describes image A. At each of the four layers we render, including L0, the readout follows image B (Figure 4b). The readout copies the tokens already present at the position rather than interpreting the context. At a visual position, the residual stream starts as the embedding of that token. Two other measurements agree. Pixel correlation between the decoded readout and the input image falls with depth, from 0.73–0.86 at L0 to 0.52–0.76 at L22 across prompts. Over the same range, the model's own generation head rises from 0.06–0.11 to 0.37–0.62. The readout's top-1 codebook entry matches the token the model emits next on at most 0.004 of positions through L14. We therefore use the rendered readout only qualitatively, and only at text token positions (Figure 4c).

At text token positions, the visual content in the subspace depends on the output modality. We fit modality-split operators on Janus-Pro in two contexts. The contexts share weights, corpus and estimator. They differ in what the model is about to emit and, by design, in their input. In both, we compute the modality index at text token positions in the middle band. This index depends on the visual readout, so we first check that the readout is reliable. At L22, it places the model's next emitted visual token in its top 20 codebook entries on 21.4% of positions, far above the chance rate of 0.12% (Appendix H.2). The excess modality index is 0.208 when the model is about to emit an image, and 0.055 when it is about to emit text. The difference is 32 times the split-half noise of 0.005 (Figure 12). We report excess over the rotated control. Appendix C.1 explains why the raw index cannot be used.

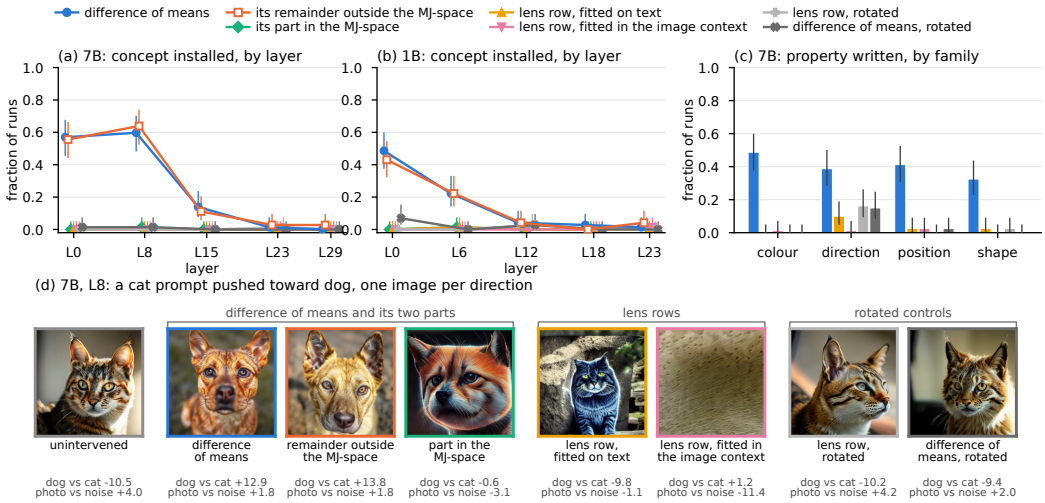

Figure 5: **A difference-of-means direction writes the object and its appearance into the image; its part in the MJ-space and the lens rows do not.** *(a, b)* Runs that install the target object, by layer, at step $2\alpha^*$. *(c)* Runs that write the target property, for the families that pass the capability gate. *(d)* One concept pair chosen by rule, with the external judge's margins below each image: the difference of means and its remainder turn the cat into a dog; its part in the MJ-space and the lens rows do not, and the rotated controls leave the cat. A run counts only if the image stays intact; intervals are Wilson 95%.

What this establishes, and what it does not. The image channel does contain visual content. How much it contains depends on what the model is about to output. But this does not show that the MJ-space is a workspace. It fails the sparsity criterion the original work (Gurnee et al., 2026) uses to identify workspace content. The content read at visual token positions only reflects the input (Appendix H).

3.5 What the image channel can be written with

This section tests the writing side of channel expressibility in the image channel. Reading shows that visual content is present in the subspace. Writing asks whether that content can be installed through the channel. Overwriting a coordinate the lens supplies does not change the image the model draws. A direction obtained without the lens does.

Lens directions do not write. In Janus-Pro-1B's image-generation context we edit prompt positions during prefill and regenerate from the same seed. We compare four conditions. The first makes no edit. The second swaps the lens coordinates of the two concepts. The third pushes along the direction between the two concepts by the same amount the swap produced. The fourth performs the same swap in a randomly rotated dictionary. The swap changes the model's margin between the host and target concepts by $+0.03 \pm 0.57$ in the text lens basis and by $+0.12 \pm 0.75$ in the visual lens basis. Both are indistinguishable from zero, and from the rotated control. The displacement-matched push moves the margin, but by damaging the image, not by installing the target (Figure 7; Appendix E.3.1). The same holds along the lens directions themselves. At Janus-Pro-7B, interventions along the rows of $W_U \operatorname{diag}(g)\, J^{T \to T}$ install the target concept in 0 of 720 runs, and along the same rows fitted in the image-generation context in 1. Lens directions are not inert. At the same norm they destroy the image far more often than a rotated direction does. But they do not write the concept.

A direction obtained without the lens writes. We therefore build the direction from a difference of means instead. At each prompt position, we subtract the average activation of prompts naming the other concepts in the same set from the activation of the prompt naming the target concept. At Janus-Pro-7B, pushing this direction into a prompt naming another concept installs the target concept in about six of every ten runs at L0 and L8, and the effect disappears beyond the middle of the network (Figure 5a). But the MJ-space is not what directly writes. We split the working direction

into a component inside the text part of the MJ-space and a remainder outside it. The MJ component carries 47% of the direction's variance yet installs the target in only 3 of 720 runs. The remainder installs it in 124, as often as the whole direction (Figure 5a, d; Table 25). Appearance is writable in the same way. On the four visual property tasks the model draws reliably, the difference of means writes the target value in 129 of 320 runs. The lens rows fitted on text and in the image-generation context succeed in only 12 and 4 (Figure 5c; protocol and per-cell results in Appendix I, colour specificity in Appendix J).

The visual content is there, and the image channel can express it, since the readout of Eq. (3) decodes to pictures. A direction obtained without the lens writes both the object the model draws and its colour, direction, position and shape. The lens's own coordinates write neither. The lens selects directions correlated with visual content, not the ones an intervention can use. On Janus-Pro, then, Jacobian-lens rows are not suitable as write directions: difference-of-means directions can install a concept, lens rows cannot. This shows that a readout lens does not automatically provide control directions.

4 Discussion

Our results separate two things that a readout lens does not automatically distinguish: what a representation can be read out as, and what it can be written with. In the cross-modal setting we study, the lens that reads does not also write: the coordinates that support readout at text positions do not control the image channel, while a direction obtained without the lens does, and the part of that direction the lens can express carries almost none of the effect. This is not a general claim about the lens. In the original text-only J-space work (Gurnee et al., 2026), a single $T \to T$ lens supports both reading and writing. Our finding differs from the text-output side only against the rotated control: on Qwen2.5-VL-7B the same coordinate swap beats a rotated control, which it does not on the image-output side, but it also fails against a displacement-matched push (Appendix E.3).

What do our experiments say about channel expressibility? In the image channel, expressibility is real but partial. Visual content is present in the subspace and its amount depends on the output modality, so the image channel can express it. But the lens's own coordinates are not the ones that address it. Expressibility in a channel is therefore not the same as control through that channel. A representation can be expressible in a channel without the readout that describes it being the handle that writes it.

This does not make the lens uninformative. It recovers a subspace with a clear causal role, and its geometry differs between modalities. What it does not give us is control. More broadly, a method that describes activations by their influence on output can recover directions that correlate with content without recovering the directions that control it. Telling the two apart requires an intervention. It also requires a positive control, to show that a null is not just a weak intervention.

We test two understanding-only models, Qwen2.5-VL-7B and Qwen3-VL-32B, and one unified model family, Janus-Pro at 1B and 7B. On Janus-Pro, the Jacobian lens, as a readout operator, does not provide control directions. This is a limitation of the reading lens, not necessarily a read–write gap of the workspace itself. Whether it persists in larger models and other unified architectures is open. Our nameability criterion is a single-token test on five synthetic classes. The ablation is confined to one band and three dictionary sizes, and the operator comparison in §3.2 has not been independently replicated. The decomposition in §3.5 uses the text dictionary, so the component we call the MJ part is the part of the MJ-space that the text dictionary spans. We focus on the image-output side. On the text-output side we ran only the coordinate swap.

A promising direction is a separate lens for writing. The MJ-lens is fitted for readout. It describes an activation by its average first-order effect on output. Nothing in that construction is designed to control the model. Unlike in the text-only case (Gurnee et al., 2026), such a lens may not write. A write lens would let us ask whether the directions that control the image channel form a coherent subspace, how it relates to the MJ-space, and whether the text and visual parts of the MJ-space are independent in both directions.

The main lesson we take from this study is that claims about a workspace should say whether they concern what can be read or what can be written, and should report the control that makes a null interpretable.

USE OF AI TOOLS

Large language models were used in preparing this submission in two roles, both auxiliary to the research. First, as a coding assistant for the experimental tooling: operator fitting, analysis and plotting scripts, and the data-handling utilities released with the paper. All numerical results were produced by executing that code on the models described, and every figure and table is generated from the stored result files rather than transcribed by hand. Second, as a writing and editing aid for prose in this document, including drafting and revision of explanatory passages. The research questions, the experimental designs, the choice of controls, the interpretation of results, and all scientific claims are the authors' own. The authors have verified every number reported here against the underlying result files and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The fitting and analysis code, the stimulus generators, and the analysis scripts that produce every table in this paper are included in the supplementary material. Operator estimation is the only expensive step and Appendix D.1 gives its cost; all downstream analyses run from the stored operator files in minutes. The estimator choices that materially change the results are enumerated in Appendix D, including two that produced incorrect results in earlier generations of this work, so that a replication can check whether it has reproduced our configuration or a different one. Stimulus sets are synthetic and generated by released code with fixed seeds; the calibration sweeps that set their difficulty levels are included, as are the per-cell clean scores, so that the floor-exclusion rule of Table 14 can be re-applied independently. For each reported comparison we release the raw per-trial outputs rather than only the aggregate, so that the confidence intervals and permutation tests can be recomputed.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*, 2021. URL https://arxiv.org/abs/2104.07143.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. URL https://arxiv.org/abs/2501.17811.

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.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, volume 15081 of *Lecture Notes in Computer Science*, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://arxiv.org/abs/2404.12390.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40056–40087, 2026. doi: 10.18653/v1/2026.acl-long.1861. URL https://aclanthology.org/2026.acl-long.1861/.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 30–45, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=94kQgWXojH.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jiang_Devils_in_Middle_Layers_of_Large_Vision-Language_Models_Interpreting_Detecting_CVPR_2025_paper.html.

Shalom Kachko, Raz Lapid, Margarita Vald, Almog Dubin, and Moshe Sipper. Through the LENS: Local geometric decomposition of vision-language model representations. *arXiv preprint arXiv:2608.00561*, 2026. URL https://arxiv.org/abs/2608.00561.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. LatentLens: Revealing highly interpretable visual tokens in LLMs. *arXiv preprint arXiv:2602.00462*, 2026. URL https://arxiv.org/abs/2602.00462.

Chungpa Lee, Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Same concept, different directions: Cross-modal feature heterogeneity in sparse autoencoders. *arXiv preprint arXiv:2606.29888*, 2026. URL https://arxiv.org/abs/2606.29888.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://openreview.net/forum?id=j84WR5ORsC.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. *arXiv preprint arXiv:2602.22766*, 2026b. URL https://arxiv.org/abs/2602.22766.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. URL https://arxiv.org/abs/1405.0312.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41157–41176, 2025. URL https://proceedings.mlr.press/v267/luo25c.html.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

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. URL https://openreview.net/forum?id=chanJGoa7f.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/89e83382abeee53b932a6df62edbf9cc-Abstract-Conference.html.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. URL https://arxiv.org/abs/2312.06681.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in VLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 9661–9675, 2025. doi: 10.18653/v1/2025.naacl-long.488. URL https://aclanthology.org/2025.naacl-long.488/.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, 2024. doi: 10.1007/978-981-96-0917-8_17. URL https://openaccess.thecvf.com/content/ACCV2024/html/Rahmanzadehgervi_Vision_language_models_are_blind_ACCV_2024_paper.html.

Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. VLMs need words: Vision language models ignore visual detail in favor of semantic anchors. *arXiv preprint arXiv:2604.02486*, 2026. URL https://arxiv.org/abs/2604.02486.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.pdf.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. URL https://arxiv.org/abs/1711.00937.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020. URL https://arxiv.org/abs/2004.12265.

Jingcheng Yang, Tianhu Xiong, Shengyi Qian, Klara Nahrstedt, and Mingyuan Wu. Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. *arXiv preprint arXiv:2602.20330*, 2026a. URL https://arxiv.org/abs/2602.20330.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Yang_Machine_Mental_Imagery_Empower_Multimodal_Reasoning_with_Latent_Visual_Tokens_CVPR_2026_paper.pdf.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 9615–9625, 2024. URL https://aclanthology.org/2024.lrec-main.840/.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *arXiv preprint arXiv:1910.07467*, 2019. URL https://arxiv.org/abs/1910.07467.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in MLLMs. *arXiv preprint arXiv:2510.24514*, 2025. URL https://arxiv.org/abs/2510.24514.

How the appendix is organised. Appendix A reviews the related work. Each of the others supplements one part of the main text and repeats none of it: Appendix B the closure argument of §2.2; Appendix C the modality index and the nameability criteria of §2.3 and §3.3; Appendix D the estimator and its cost; Appendix E the faithfulness band of §3.1 and the coordinate substitution of §3.5; Appendix F the operator geometry of §3.2; Appendix G the ablations of §3.3; Appendix H the validation of the visual readout and the output-channel comparison of §3.4 and the linearization of the visual head; Appendix I the writing protocol of §3.5, with the calibration of every lens it uses; and Appendix J the earlier writing measurements.

A Related work

Global workspace in LLMs. Gurnee et al. (2026) introduce the Jacobian lens and the J-space and establish workspace properties in text-only settings; the work has been independently replicated on open-weight models. Ours is the first extension to multimodal models—an extension the authors identify as an open question and for which they state a specific prediction. Related is work on

introspective self-report (Lindsey, 2025), which probes what models can say about their own states rather than what representations are available to be said.

Reading hidden states through the vocabulary. The Jacobian lens belongs to a family that begins with the logit lens (nostalgebraist, 2020), which decodes intermediate residual streams through the output unembedding. Geva et al. (2022) give the construction its mechanistic warrant by showing that feed-forward updates act directly in vocabulary space. The known weakness of the naive lens is that it is biased and unreliable at some depths and on some model families, which Belrose et al. (2023) address by learning a per-layer affine correction. Sparse dictionary learning is the main alternative to reading through the unembedding at all (Cunningham et al., 2023), now also in vision-language models (Pach et al., 2025); we use sparse pursuit *over* lens vectors rather than as a replacement for them, which inherits both the interpretability of a vocabulary index and the lens's failure modes. Belinkov (2022) surveys what readouts of this kind can and cannot license as claims about computation, a distinction we lean on throughout: our causal claims come from ablation, never from a readout.

When a readout is not faithful. Our faithfulness band is the sharpest form of a concern others have raised. Yom Din et al. (2024) show that a linear map from layer k to the final distribution reproduces it poorly in early layers, which is the text-only version of our measurement. Phukan et al. (2025) argue specifically that the logit lens is inadequate for VLM internal states and propose contextual-embedding readouts in its place. Our contribution here is not the observation that early-layer readouts degrade—that is established—but its magnitude and its use as an admissibility rule. On Qwen2.5-VL-7B the agreement is not merely degraded below L18, it is *exactly zero at every layer*, and the band is therefore a binary gate rather than a gradient: more than half the stack admits no readout-semantic claim at all. We also differ in what we do with it. Rather than proposing a better readout, we hold the readout fixed, measure its range, and then audit every claim in our own paper against that range (Appendix E)—including one of our own results, whose peak lies outside the faithful band and which we accordingly restate with narrower scope (Appendix F.2). Bolukbasi et al. (2021) make the complementary point that a direction can look coherently interpretable while meaning something else, which is why our binding null (Appendix H.1) is reported as a bound on the lens rather than on the network.

Interpreting visual tokens in VLMs. Visual token representations become increasingly interpretable in vocabulary space across layers, aligning with the text tokens for the image content, and object identification collapses ($> 70\%$ accuracy drop) when object-specific visual tokens are ablated (Neo et al., 2025). Architecture matters: Qwen-family models show non-zero vocabulary correspondence even in early layers, whereas LLaVA's simpler MLP adapter aligns progressively late, and a decomposition of VLM activations into local low-rank neighbourhoods (Kachko et al., 2026) similarly finds LLaVA mixing modalities late while Qwen3-VL mixes early, partially re-segregates, and recombines near the output. Yang et al. (2026a) trace multimodal circuits with transcoders and attribution graphs. These works characterize *representations*; none establishes a workspace, and none provides the functional dissociation between flexible and automatic processing that defines one. We differ in measuring **causal disposition to output** rather than current logits, and in testing **functional selectivity** via targeted ablation.

Two findings of this literature shape our method. First, **logit-lens readouts substantially understate visual-token interpretability**: Krojer et al. (2026) report 11–35% for LogitLens against a much higher rate for nearest-neighbour readout. We therefore never rely on a single readout, and we measure the range over which ours is faithful (Appendix E). Second, the literature disagrees about cross-modal sharing: task vectors appear *modality invariant* (Luo et al., 2025) and text-derived steering vectors transfer to multimodal behaviour (Gan et al., 2026), while sparse-autoencoder features for the same concept occupy *different directions* across modalities (Lee et al., 2026). Appendix F.4 tests the reconciliation that the bottleneck predicted by the workspace account implies, and finds a small effect in the predicted direction.

Nameability and visual failure. Shahgir et al. (2026) show VLM failures on fine-grained perception track nameability and that arbitrary names improve performance. We set out to supply the mechanism and the causal test, and report that the causal test does not confirm it: the behavioural regularity may be real without nameability being the gate on workspace admission (Appendix G).

The behavioural literature this sits in is by now substantial. Tong et al. (2024) identify image pairs that are visually distinct yet indistinguishable to the model, tracing the failure to the vision encoder rather than to the language model. Rahmanzadehgervi et al. (2024) find that VLMs fail at tasks as simple as counting line intersections, and frame the deficit explicitly as an inability to translate visual detail into words—a nameability claim, and therefore the closest behavioural counterpart to the hypothesis our ablations fail to support. Fu et al. (2024) assemble fourteen low-level perceptual tasks that humans solve immediately and models do not; that taxonomy is a natural stimulus set for the crossing experiment proposed in Appendix G, which our own five stimulus classes are too few to settle. Jiang et al. (2025a) and Jiang et al. (2025b) use lens-style readouts to localize and edit representations for hallucination mitigation.

Causal intervention methodology. Our ablations follow the localize-by-intervention tradition of Meng et al. (2022), and our reliance on a size- and geometry-matched control follows the recommendations of Zhang & Nanda (2024), who show that patching conclusions can reverse under different corruption schemes and metrics. Two known hazards bear directly on how we read our own results. Makelov et al. (2024) show that a subspace intervention can succeed for reasons unrelated to the hypothesized mechanism, which is why every selectivity number we report is a difference against the rotated control rather than an absolute drop. McGrath et al. (2023) show that ablated components are partly compensated for downstream, so ablation understates a component's role; this cuts against our nulls rather than our positives, and we note it where the nulls are stated. Vig et al. (2020) supply the vocabulary of direct and indirect effects that a full mediation analysis of the workspace would need, which we do not attempt here.

Latent visual reasoning. Yang et al. (2026b), Li et al. (2026a), Zhang et al. (2025) add visual latents to the decoding trajectory; Li et al. (2026b) question whether those latents are causally load-bearing. Those works add latents to the output sequence; we ask instead whether visual content already occupies the internal subspace the workspace account identifies, without adding anything to the sequence.

B WHY VERBALIZABILITY IS FIXED BY THE READOUT

This appendix gives the full argument behind the closure stated in §2.2. Section 2.2 needs only its conclusions, so we state them there and expand them here.

What the original J-lens cannot distinguish. Content may be verbal in form, or it may be non-verbal content with verbal consequences. Content is verbal in form when it is itself a word or phrase. It has verbal consequences when it changes what the model says without being made of words. Read through the vocabulary, the two look identical, because the readout reports what an activation is poised to make the model say, not what it encodes. The distinction therefore cannot be recovered from the text readout by any amount of care in reading it. It requires a readout whose symbols are not words.

Why this is not a question that can be tested. In the original J-space, a direction belongs to the space when it influences the text output. Visual content that changes what the model says therefore belongs to the space by definition. Asking whether such content has entered the space returns the definition rather than a measurement: the answer cannot fail. This is why §2.2 replaces the membership question with three questions that allow negative answers.

None of this weakens the original results on geometry, causal role, or functional behavior. In a text-only model, the claim that the space is verbalizable cannot fail, because the readout is indexed by the vocabulary and every direction is expressible in words by construction. A second output channel lets us ask a related question that can fail: can the content of the subspace also be expressed in the symbols of the image channel? Sections 3.4 and 3.5 test this question.

C METHOD DETAILS: THE MODALITY INDEX AND NAMEABILITY

C.1 THE MODALITY INDEX AND ITS CONTROLS

This section defines the modality index and the three controls it needs. The index answers a simple question: when we decompose an activation over the text dictionary and the visual dictionary together, how much of the reconstruction lands on visual atoms?

We first fix notation. Let $\mathcal{U}_{\text{text}}$ be the set of rows of the text lens matrix $W_U \operatorname{diag}(g)\, J^{m\to T}$, and let $\mathcal{U}_{\text{vis}}$ be the set of rows of the visual lens matrix $W_{\text{eff}} \operatorname{diag}(g)\, J^{m\to V}$. Here $m \in \{T, V\}$ is the modality of the source position: a text position uses $J^{T\to T}$ and $J^{T\to V}$, a visual position uses $J^{V\to T}$ and $J^{V\to V}$. The union $\mathcal{U} = \mathcal{U}_{\text{text}} \cup \mathcal{U}_{\text{vis}}$ is the combined dictionary.

An activation $h_{\ell,t}$ is decomposed by sparse non-negative pursuit over $\mathcal{U}$:

$$h_{\ell,t} \;\approx\; \sum_i \alpha_i u_i, \qquad \alpha_i \geq 0, \quad \|\alpha\|_0 \leq K. \tag{4}$$

Here α_i is a scalar coefficient and u_i is a dictionary atom. The modality index is the fraction of reconstruction mass that lands on visual atoms:

$$\mathrm{MI}(h_{\ell,t}) \;=\; \frac{\sum_{i\,:\,u_i \in \mathcal{U}_{\text{vis}}} \alpha_i \|u_i\|}{\sum_i \alpha_i \|u_i\|} \;\in\; [0, 1]. \tag{5}$$

We call this the raw modality index. The product $\alpha_i\|u_i\|$ does not change when an atom is rescaled, so the raw index cannot be moved by the two dictionaries having different norm scales. But it can be raised by three other factors even when the activation contains no visual content. Each control below rules out one of them.

C.1.1 ROTATED VISUAL DICTIONARY

The first factor is dictionary geometry. The visual dictionary may be lower-dimensional or better aligned with the dominant directions of the activation than the text dictionary, in which case a greedy selector picks visual atoms more often than their content warrants. To rule this out, we rotate the visual dictionary by right-multiplying it with a random orthogonal matrix R, so that $V^{\text{vis}} \mapsto V^{\text{vis}} R$ with $RR^\top = I$. This preserves every row norm and the Gram matrix $V^{\text{vis}}(V^{\text{vis}})^\top$ exactly, so the rotated atoms have the same lengths and the same angles among themselves as the real ones. What the rotation destroys is the relation of each atom to the residual-stream directions the model actually uses for its codebook entries.

If the index stays high after the rotation, then it is not measuring visual content. It is measuring something about the dictionary itself, which the rotation preserves. To remove that contribution, we subtract the rotated index from the raw index. We call the difference the *excess* index, and it is the statistic we report.

The control matters in practice. In the image-output context, the raw index is 0.526 and the rotated control returns 0.318, so the excess is 0.208. In the text-output context, the raw index is 0.611 and the rotated control returns 0.556, so the excess is 0.055. The two contexts are the ones of §3.4: one where the model is about to emit an image, one where it is about to emit text.

Why we do not use the raw index. The raw index, defined above as the modality index without the rotation control, gives the wrong answer. It is higher in the text-output context (0.611) than in the image-output one (0.526), even though the visual component should be larger when the model is about to draw. The reason is that the raw index is dominated by dictionary geometry, not by content. The rotated control, which destroys every correspondence between lens vectors and codebook entries, still returns 91% of the raw value in the text-output context. A statistic that a content-destroying control reproduces to within 9% is not measuring content.

The mechanism is visible in the effective dimensionality of the two readouts (Appendix H.2): the visual dictionary is much lower-dimensional (0.016–0.041) than the text one (0.29–0.64). A greedy selector prefers atoms that align with the dominant directions of the activation, regardless of what the activation encodes. Because the visual dictionary is lower-dimensional, it gets selected more often than its content warrants, and the raw index reflects that selection bias. Subtracting the rotated

control removes it. We report the excess index throughout, the same convention the paper uses for excess cross-modal consistency.

C.1.2 SIZE-MATCHED TEXT DICTIONARY

The second factor is dictionary size. A greedy pursuit selects from whichever side offers more candidate atoms more often, other things equal, and the text dictionary has $102{,}400$ atoms against $16{,}384$ visual ones. To rule this out, we repeat the decomposition with the text dictionary subsampled to $16{,}384$ atoms.

The result is that size matching raises the raw index in both contexts by about the same amount: from 0.526 to 0.546 with image output, and from 0.611 to 0.623 with text output. The difference between the contexts changes only from -0.085 to -0.077. Dictionary size therefore accounts for about 0.02 of the index in each context, and for none of the difference between them.

C.1.3 SPLIT-HALF ESTIMATE

The third factor is sampling noise. Every operator is accumulated in two halves over odd and even prompts at no extra cost, and the index is computed from each half separately. The difference between the halves is the sampling noise at the sample size in hand. We issue no verdict when that noise exceeds the difference between contexts.

The excess index is 0.211 and 0.206 in the two halves with image output, and 0.054 and 0.056 with text output. The noise level is therefore 0.005, against a between-context difference of 0.153. The difference is 32 times the noise, so it is not a sampling artifact.

C.1.4 PER-POSITION READOUT AND THE SPATIAL INDEX

The V-lens is applied independently at each position. For a position t and layer ℓ, the readout is

$$\mathrm{lens}_\ell^{\mathrm{vis}}(h_{\ell,t}) \;=\; \mathrm{softmax}\big(W_{\mathrm{eff}}\,\mathrm{norm}(J_\ell^{V\to V} h_{\ell,t})\big), \tag{6}$$

a distribution over the $16{,}384$ codebook entries. At a visual position the source modality is visual, so the operator is $J^{V\to V}$. For a visual position we take the top-ranked entry, place the 576 entries back in their positions, and pass them through the VQ decoder to obtain an image.

The operator $J_\ell^{V\to V}$ is a single $d \times d$ matrix per layer and carries no spatial index. Spatial structure in the rendered image therefore comes entirely from the spatial structure of the activations $h_{\ell,t}$, not from the operator.

The same argument explains why reading at text positions is spatially uniform. The operator $J_\ell^{T\to V}$ is defined by an expectation over all target positions t' (Eq. (2)), so it marginalizes away the target-side spatial layout. Applying it to a single text position $h_{\ell,t}$ yields one distribution over codebook entries with no information about where in the image those entries belong. A spatially structured image can only be produced by reading at many source positions at once, or by resolving the operator by target position (Appendix H.1). When those source positions are visual, the readout at each position recovers the token already present there, which is the echo reported in §3.4.

C.2 NAMEABILITY, OPERATIONALIZED

We need a definition of "nameable" because the verbalizability hypothesis turns on it. The hypothesis says that tasks whose discriminative property can be named should depend on the subspace, and tasks whose property cannot be named should not. To test this, we must decide which tasks count as nameable and which do not. We use three conditions. A task is nameable when all three hold.

1. **Single-token name.** The task's discriminative property has a name whose center word is a single token in the tokenizer vocabulary. This condition ensures that the name can be represented as one vocabulary entry, rather than as a phrase the model would have to compose.
2. **Annotator agreement.** Independent annotators give the same name for that property. This condition ensures that the name is conventional, not one person's label.

3. **Text-only sufficiency.** A text-only language model, given only the name and shared context, can solve the discrimination. This condition ensures that the name carries enough information for the task, so that a failure on the visual version cannot be blamed on an uninformative name.

The three conditions are designed to be satisfied in an obvious way by the two nameable classes and to fail in an obvious way by the three unnameable ones. We now show that this is the case.

C.2.1 SINGLE-TOKEN NAME

We checked this condition with the tokenizer of Qwen2.5-VL-7B-Instruct. The four values of the two nameable classes are single tokens with a leading space: `␣red` (2518), `␣blue` (6303), `␣left` (2115), and `␣right` (1290). Each of them is one vocabulary entry, so each can be named by a single token.

The three unnameable classes fail the condition by construction. Gabor orientation is a continuous angle, and no single token names a specific orientation. Aspect ratio is a continuous ratio, and no single token names a specific ratio. Random-polygon configuration is a combinatorial arrangement of radii, and no single token names a specific configuration. There is therefore no candidate value name to check for these three classes, which is what places them in the unnameable condition.

C.2.2 ANNOTATOR AGREEMENT

This condition is satisfied in an obvious way by the two nameable classes and fails in an obvious way by the three unnameable ones. For colour, red and blue are conventional names for the two values, and no reasonable annotator would give a different name. For arrow direction, left and right are conventional names, and no reasonable annotator would give a different name. For the three unnameable classes, there is no name to agree on: a specific Gabor angle, a specific aspect ratio, and a specific polygon configuration do not have conventional names that annotators could converge on.

We did not run a separate annotation study, because the agreement is not in doubt for either end of the contrast: the nameable classes have conventional names, and the unnameable classes have no names at all.

C.2.3 TEXT-ONLY SUFFICIENCY

This condition is also satisfied in an obvious way by the two nameable classes and fails in an obvious way by the three unnameable ones. For colour, a text-only model given "red" and "blue" as the two descriptions and asked which panel is which solves the discrimination immediately, because the names themselves state the answer. For arrow direction, a text-only model given "left" and "right" solves the discrimination immediately, for the same reason.

For the three unnameable classes, no verbal description based on the feature name can support the discrimination, because there is no feature name to begin with. One of them, aspect ratio, is a borderline case: "taller" against "wider" is a verbal description that could support the discrimination, and this is the class whose assignment would most likely change if a fuller operationalization were used.

In sum, the three conditions are satisfied in an obvious way by the two nameable classes and fail in an obvious way by the three unnameable ones, with the exception of aspect ratio noted above. The nameability contrast of §3.3 therefore rests on a definition that does not require a separate validation study at either end of the contrast.

D ESTIMATOR DETAILS AND COST

This appendix lists seven choices in the estimator that are easy to get wrong, produce no error when wrong, and materially change the reported numbers. For each choice we state what it is, why it matters, and how large its effect is on our data. We then give the cost of fitting the operators.

Corpus and sample sizes. Every number in this paper comes from these corpora and sample sizes. The fitting corpus is an interleaved image-text corpus built from COCO val2017. In an understanding

context, each prompt is an image followed by the first of its human captions. In the image-output context of Janus-Pro, the same first caption is the prompt from which the model generates an image. The image's other captions are held out and serve as ground truth in Appendix F.2.

For Janus-Pro, the operators are fitted separately in a text-output context and an image-output context. The two contexts share model weights, corpus, layer set, and estimator. They differ in the target modality and, by design, in their input. Layers are tapped at block inputs, which places our indices one below the convention that taps block outputs.

The number of prompts differs by model. We use 240 prompts for Qwen2.5-VL-7B-Instruct, 150 per output context for Janus-Pro-1B, and 60 for Qwen3-VL-32B. The Qwen3-VL-32B layers are tapped at a stride of 8, and that sample does not estimate its visual operator (Appendix F.3). Each fit accumulates its odd and even prompts into two separate estimates at no extra backward passes. Every split-half figure in this paper uses those two estimates. The lens rows written by the protocol of §3.5 are fitted on 1000 prompts per context. The full text operators its decomposition arms need are fitted on 1000 prompts (1B) and 400 prompts (7B) of pretraining text (Appendix I.3). These sizes were chosen to balance cost and split-half stability.

Why the RMSNorm gain cannot be omitted. The lens readout is $\mathrm{lens}_\ell(h) = \mathrm{softmax}(W_U\,\mathrm{norm}(J_\ell h))$. Here norm is the RMSNorm, a normalization layer that rescales the activation by its root-mean-square and then multiplies each dimension by a learned gain g. Expanding it gives

$$W_U\,\mathrm{norm}(J_\ell h) \;=\; \frac{\big(W_U\,\mathrm{diag}(g)\,J_\ell\big)h}{\mathrm{rms}(J_\ell h)}. \tag{7}$$

The scalar $1/\mathrm{rms}(J_\ell h)$ does not change which output tokens rank highest, and it does not change the shape of the coefficient distribution. Therefore the only part of the normalization that matters for the readout is the per-dimension gain $\mathrm{diag}(g)$. This matters because g is not close to constant: on Qwen2.5-VL-7B, g ranges over $[-0.059, 10.812]$, a span of about $180\times$. Dropping it is not a uniform rescaling; it changes which tokens rank highest. This is why the J-lens vectors are the rows of $W_U\,\mathrm{diag}(g)\,J_\ell$, not of $W_U J_\ell$.

Why effective dimensionality must be centered over the vocabulary. Effective dimensionality is the number of independent directions needed to carry a given fraction of the variance across the lens dictionary. To compute it, we look at how the J-lens vectors spread out. If we do not first subtract their mean, one large common direction dominates the spread. The unembedding has such a direction: nearly every vocabulary entry shares a component along it, so without centering the first principal direction is this common component, and the effective dimensionality is systematically understated. On our data, the understatement is up to an order of magnitude in early layers (0.022 vs. 0.003 at L0 at text positions) and more than half at L26 (0.711 vs. 0.316).

We compute the spread through $J^\top M J$ with

$$M = \mathrm{diag}(g)\,W_c^\top W_c\,\mathrm{diag}(g),$$

where W_c is the unembedding with its vocabulary mean subtracted. This avoids ever materializing the $n_{\text{vocab}} \times d$ matrix, which would be prohibitively large.

Why the tap must be the input to the final norm. The lens formula unembeds a specific quantity: the input to the final RMSNorm. If instead we tap the input of the last transformer block, we omit that block. The resulting J_ℓ is then not the Jacobian of the quantity the lens formula unembeds. This cannot be repaired at analysis time; it requires refitting. It matters because the last block changes the representation: the visual operator and the text operator are both substantially different when the last block is included versus omitted, with cosine similarity well below one, and occupancy, the sparsity at which the true dictionary stops beating a same-size random control, goes from censored at most positions to resolving at every position.

Why proportional truncation fails when there are few heads. The statistic is label preservation, which measures whether an attention head, when it copies a direction, copies the same direction rather than an arbitrary one. Computing it over a fixed proportion of heads, such as "the top 1% of heads by OV gain" implemented as $\max(1, \lfloor n/100 \rfloor)$, fails when the number of heads is small.

With $n = 28$ heads per layer, this expression returns exactly one head at every layer. A single head has no averaging, and the resulting numbers are dominated by idiosyncrasies of that head; the rotated control, which should be near zero by construction, reached -0.318 on individual layers. We therefore use the top decile with a floor of four heads, report the all-head median as the primary figure, and attach bootstrap intervals. Any proportional truncation should be checked for how many items it actually selects at the n in hand.

Why binarized readouts are swamped by response bias. The problem is that a two-alternative forced-choice measure scored as a binarized decision, $\mathbb{1}[\,\mathrm{logit}(\text{“ same”}) > \mathrm{logit}(\text{“ different”})\,]$, is dominated by response bias. On synthetic stimuli, the model has a strong “same” bias: most cells answered “same” on 12 of 12 trials, with margins of $+2$ to $+8$ that rarely change sign, while the stimulus set is balanced 12/12. Accuracy is then pinned at 0.500 by construction, and no ablation effect can be detected no matter how large it is. We therefore score with the continuous margin and report AUC. A constant bias shifts all margins by the same amount, so it does not change the ranking of same- and different-trials, and AUC is unaffected by it.

Changepoint estimation. Onset layers in Appendix F.2 are estimated two ways: a three-segment least-squares fit and a departure statistic. The two disagree on some curves. We explain why in Appendix F.2, where Table 5 reports both.

Why the source modality of the operator must match the positions read out. A modality-split operator $J_\ell^{m\to m'}$ averages over source positions of modality m. It may therefore only be used to read out activations at positions of that same modality. The rule is about the source modality m, not about requiring $m = m'$. Both cross-modal operators are legitimate: $J^{T\to V}$ is used at text positions, and $J^{V\to T}$ is used at visual positions.

Concretely: at text positions the two available operators are $J^{T\to T}$ and $J^{T\to V}$; at visual positions they are $J^{V\to T}$ and $J^{V\to V}$. We use $J^{V\to T}$ at visual positions in the ablation of §3.3 (Appendix G.1), and $J^{T\to V}$ at text positions in the modality index of §3.4.

The failure mode is a source-modality mismatch. Applying $J^{T\to V}$—source positions text—to activations at visual positions gives a top-20 next-token hit rate of 0.000 at every layer, which is what a broken readout looks like. With the source modality matched ($J^{V\to V}$), the same check gives 0.214 at L22 on forty rollouts, rising monotonically with depth (Appendix H.2). Nothing errors, and the mismatched numbers are plausible-looking zeros. Any readout check reporting exactly chance should be suspected of a source-modality mismatch before it is believed.

Why L0 is excluded. L0 is excluded from all changepoint fits. Its tap is the embedding output, where the Jacobian is near-identity and $\|J_0^{T\to T}\| = 23193$. That is roughly $300\times$ every other layer and four orders of magnitude above the visual operator at the same layer (2.15). A layer whose operator norm is four orders of magnitude larger than the others is not comparable with the rest of the curve.

D.1 How much computation fitting the operators requires

Computing $\partial h_{\text{final}}/\partial h_\ell$ exactly requires d_{model} backward passes, or the same number of forward-mode Jacobian-vector products. The reference implementation reports that lens quality saturates quickly and that about 100 prompts suffice, against the 1000×128 tokens used in the original paper.

We measured how the modality split in Eq. (2) changes this cost. Three facts, each verified empirically on a small causal network:

1. **Depth is free.** A single backward pass yields gradients at every layer simultaneously. The number of tapped layers therefore does not enter the cost—only memory for the accumulators.
2. **The source split is free.** One backward pass yields gradients at all source positions t. Bucketing them by $\mu(t)$ separates $J^{T\to T}$ from $J^{V\to T}$ at no extra cost.
3. **The target split is not free.** Which positions carry the cotangent determines the set summed over. Therefore $J^{T\to T}$ and $J^{V\to V}$ require separate backward passes.

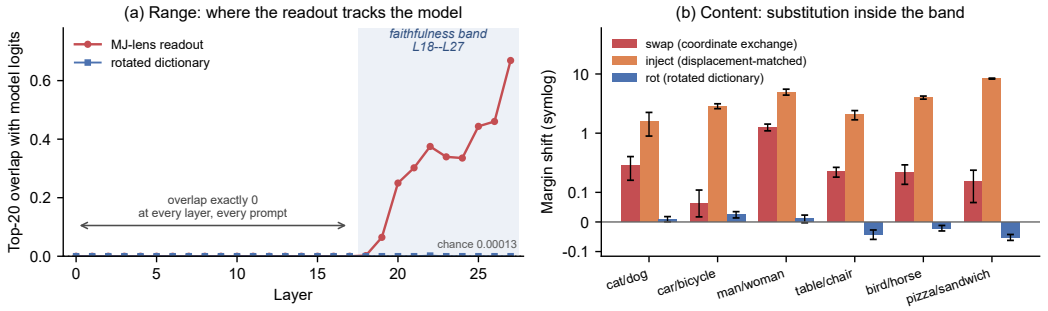

Figure 6: **The readout tracks the model's output only from L18, and inside that band a lens coordinate is not what the model computes with.** *(a)* Top-20 overlap between the $J^{T\to T}$ readout and the model's own final-layer logits (24 prompts), against a rotated dictionary. Below L18 the overlap is exactly zero at every layer and for every prompt, which turns a gradient into an admissibility rule; L18–L27 is the faithfulness band. *(b)* Coordinate substitution inside the band (40 images per pair): the margin shift from exchanging two concepts' lens coordinates (swap), from a push of the same displacement (inject), and from the swap in a rotated dictionary (rot). Pooled over pairs, swap shifts the margin by $+0.37$ and inject by $+3.97$: the swap clears the rotated control on every pair and loses to the displacement-matched push on every pair.

The cost is therefore

$$\#\text{backward} \;=\; d_{\text{model}} \times |\{\text{target modalities}\}| \times \#\text{prompts}, \tag{8}$$

i.e. $2\times$ the text-only cost when visual targets are needed. Note that understanding-only VLMs have no visual output. Therefore the layer-structure and operator comparisons of Appendix F require only text-target operators and cost exactly as much as a text-only replication. The $2\times$ applies only to the output-context comparison of Appendix H.2.

Rows instead of operators. An intervention uses a few dozen lens rows, not the full $d\times d$ operator. Placing a row's cotangent $g\odot W_U[w]$ on every target position and back-propagating once per prompt gives that row at every layer from a single backward pass. The cost of calibrating K rows is therefore $\lceil K/k\rceil$ batched backward passes per prompt, for a batch of k cotangents, instead of d passes. Here d is the model dimension. This is what made $n = 1000$ prompts affordable in every context of the calibration in Appendix I.3: about 5 s per prompt on Qwen2.5-VL-7B and 8 s on Qwen3-VL-32B in the understanding context, against one minute per prompt for the full $J^{T\to T}$ of Janus-Pro-7B.

E WHERE THE READOUT IS VALID, AND WHAT IT CAN SUPPORT

Section 3.1 introduces two checks but gives only their conclusions. This appendix explains how each check is performed. The **range check** determines the depths at which the lens readout agrees with the model's own output. We call this depth range the **faithfulness band**, the range in which the readout has a measurable correspondence to the model's final output. The **content check** determines whether a particular reading faithfully reflects what the model computes with. We give the per-layer values behind the range check, the same check on the unified model, an audit of our own claims against the faithfulness band, and the content check in full.

E.1 RANGE CHECK: WHERE THE READOUT IS VALID

We ask: at which depths does the lens readout agree with the model's own output? The higher the agreement, the more trustworthy the readout.

Design. We read out at the final text position. We choose that position because $J^{T\to T}$ ought to agree best there: h_{final} is taken from that position, so the operator and the unembedding describe the same quantity. At that position we intersect the top 20 of $\text{lens}_\ell(h)$ with the top 20 of the model's own final-layer logits. We do this over 24 prompts on Qwen2.5-VL-7B-Instruct. The chance rate is 0.00013.

Layer	L0–L17	L18	L19	L20	L21	L22	L23	L24	L25	L26	L27
Overlap	0.000	0.002	0.065	0.250	0.302	0.375	0.340	0.335	0.444	0.460	0.669
rot	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000

Result: the readout is zero below the faithfulness band and rises above it. The overlap is exactly zero at every layer from L0 to L17, on every one of the 24 prompts. It is not zero on average; it is zero on each prompt individually. It becomes nonzero at L18 and rises from there. On this model, the faithfulness band is therefore L18–L27.

Why is this credible? The rotated control preserves row norms and the Gram spectrum exactly, and destroys only the correspondence to vocabulary entries. If the signal came from the unembedding itself, the rotated control would give the same result. It does not, so the signal between L18 and L27 is carried by the operator.

That early-layer readouts degrade is known (Yom Din et al., 2024; Phukan et al., 2025). Logit-lens readouts are also known to understate the interpretability of visual tokens (Krojer et al., 2026). What is specific here is that the agreement is identically zero over eighteen consecutive layers. This tells us that below the faithfulness band, a zero result from the readout cannot distinguish "no effect" from "no range in the instrument." Every readout-semantic claim below therefore has to state which side of the faithfulness band it falls on.

The same check on the unified model. On Janus-Pro-1B we score the text J-lens of the $n = 1000$ fit (Appendix I.3) against the model's own output at every position of held-out pretraining text. The operator from one half of the corpus reads the texts of the other half. This protocol differs from the one above in position set and model, so the two are not directly comparable. The shape is still informative. Overlap with the model's top 20 is 0.018 at L0, 0.081 at L12, 0.200 at L18, 0.373 at L22, and 0.558 at L23. It is never exactly zero. The rotated control is 0.000–0.001 throughout. The per-layer next-token hit rates of the same run are in Table 21, where they serve as the reference for the linearization of the visual head.

E.2 Audit: whether our own claims are within range

Once the range check is established, we audit our own claims. Some measurements look only at the geometry or statistics of activations, and never at which tokens the readout returns. Other measurements depend directly on which tokens the readout returns. The first kind is not bounded by the faithfulness band. The second kind is.

Claims not bounded by the faithfulness band. The operator geometry, effective dimensionality, and CKA of §3.2 belong to the first kind. They do not look at tokens, only at numbers, and therefore hold at every layer regardless of the faithfulness band.

Claims below the faithfulness band. Two measurements that depend on readout tokens sit below the band. We state their range in place. The caption-hit contrast (Appendix F.2) and the rendered readout (§3.4) both peak below L18, where the lens has zero overlap with the model's own output.

Claims that need extra care. The ablation experiments of §3.3 intervene in L9–L17. Their dependent variable is model behaviour and does not need readout tokens. But their target is chosen by the lens's token ranking in exactly the layers where that ranking has no measured relation to the output. These experiments therefore establish a fact about the subspace the lens selects—well defined, and beating a Gram-matched control—but they do not settle whether that subspace is the workspace.

E.3 Content check: coordinate substitution

We ask whether the coordinates the lens reads out faithfully reflect the model's computation. If they do, then replacing one concept's lens coordinates with another's should change the model's behaviour accordingly.

Design. For a concept pair (s, t) we take the two lens vectors $V = [v_s\ v_t]$. We read the activation's coordinates in that basis by pseudo-inverse. We write back the coordinate-swapped activation $h' = h + V(\sigma(c) - c)$, leaving the component in the orthogonal complement untouched. The dependent variable is the change in the two-alternative margin $\text{logit}_t - \text{logit}_s$ on a forced-choice question about the image. This measure is immune to answer bias because a bias shifts both terms equally.

Two controls are required and both are reported. The first, `inject`, pushes along $(v_t - v_s)$ by a step whose norm in activation space equals the displacement the swap actually produced. The second, `rot`, performs the same swap on a dictionary right-multiplied by a random orthogonal matrix. The criterion is that the swap must exceed both controls.

Results. Six pairs (`cat`/`dog`, `car`/`bicycle`, `man`/`woman`, `table`/`chair`, `bird`/`horse`, `pizza`/`sandwich`). Substitution is applied at layers L8–L20, with 24–40 images per pair (Table 1).

Table 1: Mean margin shift across the six pairs, ± standard error. `sel` restricts to the images with the largest clean |margin|. Displacement norms of the swap and inject arms agree to four decimal places within every pair, so the comparison is not confounded by intervention size.

Run	α	n/pair	swap	inject	rot
`faith_sub`	1	40	$+0.077 \pm 0.150$	—	-0.008 ± 0.013
`faith_sub_a1`	1	24	$+0.042 \pm 0.169$	$+2.284 \pm 1.646$	-0.004 ± 0.013
`faith_sub_a4`	4	24	-0.716 ± 1.441	$+3.565 \pm 1.619$	-0.028 ± 0.063
`faith_sub_a16`	16	24	-0.338 ± 1.661	—	$+0.380 \pm 0.135$
`faith_sub_sel`	1	24	$+0.728 \pm 0.283$	$+4.132 \pm 0.913$	-0.012 ± 0.005
Restricted to the faithfulness band, L18–L27					
`faith_sub_inband`	1	40	$\mathbf{+0.365 \pm 0.181}$	$\mathbf{+3.966 \pm 1.015}$	-0.012 ± 0.013

Table 2: Per-pair breakdown of the in-band run (L18–L27, n=40). Unlike the out-of-band runs, every pair moves in the same direction and every pair separates from its rotated control.

Pair	swap	inject	rot
`cat/dog`	$+0.281 \pm 0.121$	$+1.568 \pm 0.674$	$+0.009 \pm 0.009$
`car/bicycle`	$+0.063 \pm 0.046$	$+2.866 \pm 0.272$	$+0.024 \pm 0.011$
`man/woman`	$+1.259 \pm 0.164$	$+4.967 \pm 0.570$	$+0.010 \pm 0.013$
`table/chair`	$+0.223 \pm 0.042$	$+2.045 \pm 0.363$	-0.043 ± 0.016
`bird/horse`	$+0.214 \pm 0.077$	$+3.997 \pm 0.239$	-0.021 ± 0.009
`pizza/sandwich`	$+0.151 \pm 0.086$	$+8.354 \pm 0.160$	-0.052 ± 0.010

Reading. Against the rotated control, the swap is decisive in the highest-powered run: $+0.728$ against -0.012, with non-overlapping intervals. The dictionary's correspondence to vocabulary entries is therefore doing work. The readout is not pure confabulation.

Against the displacement-matched control it fails, and the gap is large. For the same activation-space step, the `inject` arm moves the margin 5.7 times as far as the swap. In three of the five runs, the swap's own mean does not exclude zero. At $\alpha = 4$ it reverses sign. By the stated criterion, the protocol item does not pass.

The faithfulness band does not explain it. The first five runs were applied over L8–L20, of which ten layers lie below the faithfulness band. There, the readout's top-20 has exactly zero overlap with the model's own output. The natural reading was that a coordinate carrying no measurable relation to the output cannot be swapped to any effect, and that the failure was an artefact of the depth range rather than a property of the readout. We tested that reading by repeating the experiment inside the band. It does not survive.

Restricting to L18–L27 changes the result in one respect, not in the one that matters. Consistency improves markedly: all six pairs now move in the same direction, and each separates from its own rotated control (Table 2). The out-of-band runs had mixed signs. The magnitude does not improve. It falls, from $+0.728$ to $+0.365$, while the displacement-matched control is unchanged. The ratio of

swap to inject therefore drops from roughly one in six to one in eleven. Whatever the swap is failing to do, it fails to do it just as much where the readout demonstrably tracks the output.

Conclusion. We report the content check as not discharged, and we withdraw the range-based explanation. What survives is narrower and better supported than what we first wrote. The lens dictionary carries real correspondence to the model's computation, since swapping along it beats a geometry-matched rotation consistently and inside the verified range. But a coordinate swap is a far weaker intervention on behaviour than an equally-sized generic push. That means the readout's coordinates are not the directions the model's computation is most sensitive to. Why a matched push along $(v_t - v_s)$ should dominate the principled swap is not something these data settle, and we flag it as the open question this section leaves.

The parts of the content check we have not run. The reference protocol has three further items beyond substitution, none of which we have executed. The first is a retention check: confirming readout content that appears in neither prompt nor output as a task-relevant intermediate, with an independent linear probe and with human annotation. The second is agreement between the J-lens, a nearest-neighbour readout (Krojer et al., 2026), and the rendered visual readout on the same activations, including the rate at which each resists interpretation. The third is an interpretability rate benchmarked against the published logit-lens figures of 11–35% under a matched protocol. The content check is therefore reported as not discharged on the one item we ran and as not attempted on the other three.

E.3.1 The same check in the image channel

The range check and the content check were both run on the text readout. We now move the content check to the image channel of Janus-Pro-1B, to see whether coordinate substitution also fails on the image-output side.

Design. In the image-generation context of Janus-Pro-1B, a text prompt is followed by 576 emitted visual tokens. The prompt names the host concept s. The edit is applied at the prompt positions during prefill, at L8, L10, L12, and L14, to the conditional half of the classifier-free-guidance batch (scale 5). It reaches every emitted token through the key–value cache. Generation is re-run from the same seed, and the four conditions are decoded together by the model's own VQ decoder. The conditions and the patch arithmetic are those of the check above, at $\alpha = 1$: `clean`; `swap`; `inject`, along $v_t - v_s$ with the displacement the swap produced; and `rot`, the swap on the dictionary right-multiplied by a random orthogonal matrix.

The dependent variable is the model's own two-alternative margin $\mathrm{logit}_t - \mathrm{logit}_s$ on its own output image. The question is asked as "Answer with exactly one word. Does this image contain a $\{s\}$ or a $\{t\}$? Answer:", and reported as the shift from `clean`. Since the prompt names s, a swap that installs t raises it.

Each concept has a direction in each dictionary. The text direction is the word's row of $W_U \,\mathrm{diag}(g)\, J^{T\to T}$. The visual direction is the mean of the $W_{\text{eff}} \,\mathrm{diag}(g)\, J^{T\to V}$ rows of the 16 codebook entries whose readout at the prompt naming the concept most exceeds the mean readout at the prompts naming the other seven concepts. Both operators are fitted in the image-generation context.

The text directions of a pair are far more collinear than the visual ones. The median Gram condition number over the 16 pair–layer cells is 155 (range 56–390) for text against 3.2 (range 1.1–31) for visual. This makes the text arm the weaker measurement. A layer above 400 would have been skipped, and none was. Four pairs, *dog*→*cat*, *car*→*bicycle*, *pizza*→*sandwich*, and *bird*→*horse*, are each run on eight scenes.

Results. Averaged over the four pairs, the swap moves the margin by $+0.03 \pm 0.57$ with the text dictionary and $+0.12 \pm 0.75$ with the visual one. Both are indistinguishable from zero and from the rotated control (-0.45 ± 0.84 and $+0.40 \pm 0.39$). The displacement-matched push moves it by $+5.13 \pm 1.86$ and $+4.00 \pm 2.61$ (Table 3). The swap and the push displace the activation by the same amount in every pair, between 0.038 and 0.069 of its norm. The gap between them is therefore not a difference in how hard we pushed. The rotated swap displaces it by less, 0.031–0.039.

Table 3: Coordinate substitution in the image channel of Janus-Pro-1B, four pairs, eight scenes each. Margin shift: change of the model's own two-alternative margin from the unintervened image, mean ± standard error over scenes (over pairs in the last two rows). Pixel difference: mean absolute difference from the unintervened image on a 0–255 scale. Tokens kept: the fraction of the 576 emitted tokens equal to the unintervened run's at the same position.

		margin shift			pixel difference			tokens kept		
pair	dictionary	swap	inject	rot	swap	inject	rot	swap	inject	rot
dog→cat	text	$+0.23 \pm 0.16$	$+7.23 \pm 0.18$	$+0.00 \pm 0.25$	31.5	53.4	33.8	0.057	0.001	0.030
dog→cat	visual	$+0.53 \pm 0.16$	$+0.03 \pm 0.19$	$+1.13 \pm 0.73$	52.2	62.8	65.1	0.013	0.003	0.006
car→bicycle	text	$+1.31 \pm 1.13$	$+9.03 \pm 0.42$	$+0.84 \pm 0.65$	70.9	83.8	61.2	0.015	0.002	0.016
car→bicycle	visual	$+0.91 \pm 0.49$	$+9.53 \pm 0.35$	$+1.00 \pm 0.66$	40.9	72.9	40.8	0.009	0.001	0.010
pizza→sandwich	text	-1.45 ± 1.74	$+0.78 \pm 1.34$	-2.91 ± 1.15	39.9	63.6	46.1	0.045	0.004	0.019
pizza→sandwich	visual	$+1.13 \pm 1.18$	-0.91 ± 1.55	-0.45 ± 1.16	45.6	76.0	51.3	0.035	0.002	0.021
bird→horse	text	$+0.03 \pm 0.36$	$+3.45 \pm 2.33$	$+0.27 \pm 1.14$	51.0	80.5	54.3	0.031	0.001	0.009
bird→horse	visual	-2.09 ± 1.41	$+7.33 \pm 1.33$	-0.08 ± 2.08	53.8	83.4	44.8	0.006	0.001	0.006
mean over pairs	text	$+0.03 \pm 0.57$	$+5.13 \pm 1.86$	-0.45 ± 0.84	48.3	70.3	48.9	0.037	0.002	0.018
mean over pairs	visual	$+0.12 \pm 0.75$	$+4.00 \pm 2.61$	$+0.40 \pm 0.39$	48.1	73.8	50.5	0.016	0.002	0.011

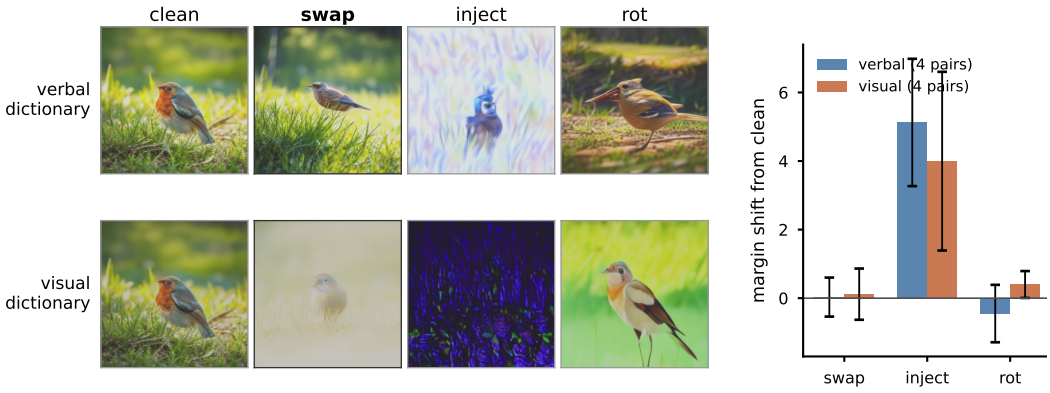

Figure 7: **Exchanging two concepts' lens coordinates leaves the image unchanged; only the displacement-matched push moves the margin, by damaging the image.** Janus-Pro-1B, image-generation context, prompt *a bird on the grass in daylight*, target concept *horse*. The four arms share a seed and differ only by the edit, which is applied at prompt positions during prefill and reaches all 576 emitted tokens through the key–value cache. *Right:* the shift in the model's own two-alternative margin, averaged over concept pairs. A coordinate the readout had correctly identified would put `swap` above both controls; instead `swap` sits at zero, and the displacement-matched `inject`—the only arm whose margin moves—does so by leaving the image distribution rather than by drawing a horse.

What the push does to the picture. At the same displacement, the push changes the picture more than the swap does. The mean absolute pixel difference from the unintervened image is 70.3 and 73.8 for the two dictionaries, against 48.3 and 48.1 for the swap and 48.9 and 50.5 for the rotated control. It also keeps almost none of the unintervened image's tokens, at most 0.4% of them in any pair, against up to 5.7% for the swap and 3.0% for the rotated control. That the pushed pictures leave the distribution of the model's images rather than depict the target is judged from the images (Figure 7). No distributional statistic was computed.

Replication with operators fitted in the other context. An earlier run with the same pairs, scenes, layers, and seed used operators fitted in the text-output context. It reaches the same conclusion. The swap moves the margin by -0.48 ± 0.81 (text) and -0.38 ± 0.53 (visual). The rotated control moves it by $+0.60 \pm 1.02$ and $+0.55 \pm 1.39$. The push moves it by $+3.70 \pm 2.62$ and $+4.21 \pm 2.59$, with pixel differences of 67.0 and 73.2 for the push against 46.8–50.5 for the other two conditions. The null therefore survives changing the operator it depends on.

Table 4: Structural signatures reused from Gurnee et al. (2026). All are computable without interpreting readout semantics.

Signature	Definition	Workspace behavior
CKA block structure	Centered kernel alignment (Kornblith et al., 2019) between layers' J-lens Gram matrices	Three blocks: sensory / workspace / motor
Excess kurtosis	Excess kurtosis of the lens logit distribution	$\approx$ 0 early; rises at onset; falls at motor transition
Positional autocorrelation	Persistence of top-1 lens token from t to $t+\Delta$ vs. a position-shuffled null	Near null early; peaks across the middle band
Effective dimensionality	Fraction of residual dimensions capturing a given share of variance across $W_U J_\ell$	Small early; rises sharply at onset
Occupancy K^*	K at which marginal reconstruction gain falls below a size-matched random control	$\approx$ 0 early; plateaus $\approx$ 25 across the band
MLP gain	Output norm of MLP block $\ell{+}1$ applied to a unit direction, normalized by median over isotropic random directions	$\approx$ 1 outside the band; $\approx$ $10\times$ within
OV gain / label preservation	Attention-head broadcast metrics on W_{OV}	Identifies a small set of broadcast heads

The coordinates being exchanged. In the visual dictionary, the coordinates do not encode which concept the prompt names. Over the 16 pair–layer cells, the coordinate of the concept the prompt names exceeds the other concept's in 77% of the text cells (mean difference $+0.171$) but in only 31% of the visual ones (mean -0.073). On the visual side, a swap therefore exchanges two values that do not distinguish the concepts to begin with.

F Where is the multimodal workspace?

This appendix supports Section 3.2. We describe the structural signatures we borrow from the original work, the layer profiles at visual and text positions, the comparison between the visual-to-text and text-to-text operators, the broadcast measurements, the response to a visual morph, and the occupancy and capacity measurements. For each part, we say what question it answers and why the measurement suits that question.

F.1 Structural signatures

The original work marks out a layer band using signatures that do not depend on whether the readouts make sense to a human reader. We reuse all of them. This matters because our claims about where the multimodal workspace sits should not rest on our own reading of the readouts. Table 4 lists the signatures and what each is expected to look like inside the band.

A few of these signatures need a word of explanation. CKA (centered kernel alignment) measures how similar two linear maps are: 0 means unrelated, 1 means identical. Excess kurtosis measures how heavy the tails of a distribution are; a normal distribution scores 0, and a positive value means heavier tails, that is, a sparser and more concept-like distribution. Effective dimensionality is the number of independent directions needed to carry a given share of the variance. Occupancy is the sparsity at which the true dictionary stops beating a same-size random control. MLP gain and OV gain measure how much a layer or an attention head amplifies a direction, compared with random directions. Label preservation measures whether a head, when it copies a direction, copies the same direction rather than an arbitrary one.

F.2 Layer structure at visual vs. textual positions

We ask whether the visual subspace is built up and then spent, as Section 3.2 claims. That claim depends on the shape of the effective-dimensionality curve. This section gives that curve and the other measurements that support it.

Design. We fit $J^{T\to T}$ and $J^{V\to T}$ on Qwen2.5-VL-7B-Instruct (240 prompts, all 28 layers) and on Qwen3-VL-32B (Bai et al., 2025a) (60 prompts, every eighth layer) over the corpus of Appendix D. We compute all signatures of Table 4, separately for image-patch positions and text positions.

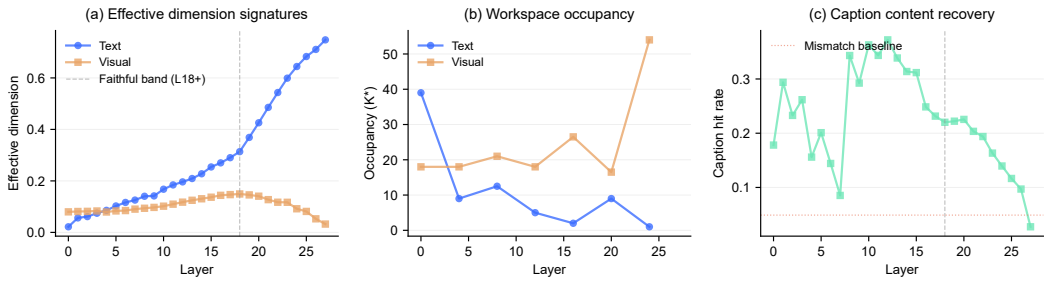

Figure 8: **The lens dictionary spans the most directions mid-network at visual positions and keeps expanding with depth at text positions** (Qwen2.5-VL-7B, 240 prompts). (a) Effective dimensionality rises monotonically at text positions (0.022 at L0 to 0.748 at L27) but is non-monotone at visual positions, peaking at 0.149 (L18) and falling back to 0.053 by L26. This signature interprets no readout semantics, which is why we rest the modality dissociation on it. (b) Occupancy K^* against the rotated control, at the seven layers on which it was estimated. Visual exceeds text at every layer but L0 and the gap is largest at L24 (54 vs. 1). The text-side collapse toward 1 appears against this control only (Appendix F.6) and is not a replication of the plateau reported by Gurnee et al. (2026). (c) The $K = 25$ sparse decomposition at visual positions recovers content words from the image's *other* human captions, peaking at 0.372 (L12) against a mismatched-caption baseline of 0.049 (dotted). Dashed vertical line marks the onset of the faithfulness band (Appendix E). Panel (c) peaks well below that band, so it licenses a claim about content correspondence only, not about what the model is preparing to say.

Table 5: Workspace onset by position modality. Two complementary estimators, both bootstrapped over prompts: *3-seg* fits two changepoints by least squares; *depart* is the first layer that persistently exceeds the early plateau by 10% of the signal's dynamic range. The two disagree by design—on a "0 → small → large" curve the 3-segment fit places both changepoints at the terminal motor rise and misses the true onset (see top-k accuracy, text row). L0 is excluded: its tap is the embedding output, where the Jacobian is near-identity and $\|J^{T\to T}\| = 23193$, four orders of magnitude above the visual operator at the same layer. Qwen3-VL-32B was fit at stride 8 (60 prompts), so its layer resolution is 8× coarser and K^* is reported at its own mid-band.

		Top-k acc.		**Caption hit**		**Eff. dim.**	K^*
Model	**Positions**	3-seg	depart	3-seg	depart	depart	(mid-band)
Qwen2.5-VL-7B (28L)	text	L15	L3	L23	L10	L10	5–2
	visual	L8	L6	L8	L4	L7	18–26
Qwen3-VL-32B (64L)	text	L32	L16	L32	L16	L40	7–12
	visual	L32	L32	L32	L32	L32	3–11

Why we expect a difference. The original work could not tell whether the early-layer absence of J-space content is a fact about the model or a failure of the lens. Multimodality gives us a way to separate the two. If early layers really are a sensory stage, the sensory band should last longer at visual positions, because visual input starts further from the vocabulary and needs more depth to be put into words.

What we predicted. Onset of the workspace is later at visual positions than at text positions, and the two converge in the middle band. If the onset layers were the same, that would mean the sensory/workspace boundary is a property of network depth, not of input format.

Result: the prediction fails in the direction we stated. On the 7B model, visual positions leave baseline earlier, not later, than text positions. The visual signatures put onset at L4–L8, while the text caption-hit signature does not reach its changepoint until L23. The excess-kurtosis signature resolves cleanly at visual positions (onset L7, bootstrap interval $[7, 7]$, $\mathrm{snr} = |\mathrm{step}|/\mathrm{residual\ SD} = 4.49$), but two other cells do not resolve (text caption-hit, both autocorrelation rows), and we report those as unresolved rather than as nulls.

This reverses the reading we proposed above, and the reinterpretation is simple. The late departure at text positions is not a late workspace onset but the terminal motor rise: at text positions, the lens

Table 6: Effective dimensionality of the lens dictionary at every fitted layer: the fraction of residual dimensions carrying 90% of the variance of the J-lens vectors, at text and at visual source positions.

Qwen2.5-VL-7B-Instruct

Layer	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
text	0.022	0.056	0.061	0.074	0.086	0.103	0.116	0.126	0.140	0.142	0.168	0.184	0.196	0.209
visual	0.080	0.081	0.082	0.083	0.079	0.083	0.085	0.090	0.094	0.097	0.102	0.110	0.117	0.125

Layer	L14	L15	L16	L17	L18	L19	L20	L21	L22	L23	L24	L25	L26	L27
text	0.228	0.254	0.270	0.290	0.314	0.369	0.426	0.485	0.543	0.599	0.644	0.683	0.711	0.748
visual	0.131	0.137	0.144	0.147	0.149	0.146	0.141	0.127	0.117	0.117	0.092	0.081	0.053	0.032

Qwen3-VL-32B, every eighth layer

Layer	L0	L8	L16	L24	L32	L40	L48	L56
text	0.010	0.034	0.054	0.071	0.089	0.122	0.325	0.593
visual	0.017	0.031	0.031	0.040	0.061	0.085	0.142	0.213

converges onto the model's own next-token prediction (top-20 accuracy rises from 0.000 to 0.901 across depth). Visual positions never undergo that convergence (top-20 accuracy stays below 0.10 at every layer). So their earlier departure reflects entry into the workspace, not preparation of the output. Under this reading, the two modalities are being measured at opposite ends of the pipeline, and the onset comparison as we first framed it does not isolate what we wanted. We report it as a clarification of the estimator, not as evidence for or against the bottleneck.

On Qwen3-VL-32B the ordering is different again: visual onset is at L32 for every signature while text departs at L16. At stride 8, the layer resolution is too coarse to tell L16 from L32 with confidence, and the visual operator of that fit is well estimated only at L48 and L56 (Appendix F.3), so we draw no conclusion from the reversal.

Result: effective dimensionality separates the two modalities. Effective dimensionality is the number of independent directions needed to carry 90% of the variance across the lens dictionary. At text positions it rises smoothly with depth, from 0.022 at L0 to 0.748 at L27. At visual positions it is not monotone: it peaks at 0.149 around L18 and falls back to 0.053 by L26 (Figure 8a, Table 6). This rise-then-fall is what we expect if visual content enters a shared workspace in the middle of the network and is then consumed or translated, rather than being carried all the way to the output. On Qwen3-VL-32B only the first half appears: text reaches 0.593 at L56 against 0.213 at visual positions, but the visual curve rises at every sampled layer up to L56, so the fall-back is not seen there. The visual operator of that fit is well estimated only at L48 and L56, so its values below L48 carry sampling noise of unknown size.

Result: sparse decompositions at visual positions carry the image's caption content. We use the other human captions of the same image as ground truth. The prompt contains only the first caption, and causal masking stops visual positions from attending to any of the others. At visual positions, the $K = 25$ sparse non-negative decomposition picks caption content words at 10–24$\times$ the rate of a mismatched caption, peaking at L12 (0.372 vs. 0.049; Figure 8c). The absolute hit rate at visual positions exceeds that at text positions at every layer before L20. The individual tokens recovered are partly interpretable at this corpus size, so we rest the claim on the matched-vs-mismatched contrast rather than on reading individual tokens.

Range. This measurement sits below the faithfulness band of Appendix E: at L12, the top-20 of the lens has zero overlap with the model's own final-layer logits. The contrast is not thereby invalidated, because its ground truth is the image's other human captions and its control is a mismatched caption, and neither refers to the model's output distribution. But the range of the conclusion is fixed by it: visual positions carry content that corresponds to the image, but they are not thereby preparing to say that content. The claim that the model uses this content comes from the causal ablation of Appendix G.1, which intervenes on activations and scores generated text, not from the readout.

F.3 Is $J^{V\to T}$ the same operator as $J^{T\to T}$?

We ask how different the two operators are. Section 3.2 claims that splitting by modality is necessary, which requires the two operators not to coincide. This section measures how far apart they are.

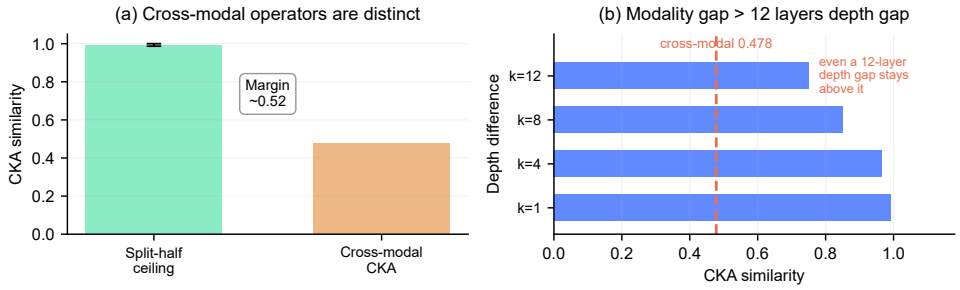

Figure 9: **The visual-to-text and text-to-text operators differ by more than twelve layers of depth do within one modality.** (a) Cross-modal CKA in the middle band averages 0.478, against a split-half ceiling of 0.987–1.000 obtained by estimating the same operator twice from disjoint halves of the prompts. The ceiling is what makes the comparison interpretable: a bare similarity number would not be. (b) The same difference expressed in units of network depth. Walking k layers forward along the text operator gives CKA 0.991 (k=1), 0.966 (k=4), 0.851 (k=8) and 0.751 (k=12), all above the cross-modal 0.478, so changing modality moves the operator further than twelve layers of depth do within one modality.

Table 7: Cross-modal CKA and the split-half ceilings of the two operators at every layer of Qwen2.5-VL-7B-Instruct (240 prompts, halves of 120).

Layer	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
$\mathrm{CKA}(J^{V\to T}, J^{T\to T})$	0.387	0.666	0.665	0.682	0.647	0.657	0.660	0.675	0.588	0.575	0.497	0.474	0.492	0.488
ceiling, $J^{T\to T}$	1.000	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999	0.998	1.000	1.000	1.000	1.000
ceiling, $J^{V\to T}$	0.987	0.989	0.989	0.989	0.990	0.988	0.988	0.990	0.990	0.991	0.991	0.990	0.990	0.991
Layer	**L14**	**L15**	**L16**	**L17**	**L18**	**L19**	**L20**	**L21**	**L22**	**L23**	**L24**	**L25**	**L26**	**L27**
$\mathrm{CKA}(J^{V\to T}, J^{T\to T})$	0.476	0.453	0.445	0.432	0.452	0.382	0.394	0.389	0.451	0.452	0.424	0.436	0.407	0.418
ceiling, $J^{T\to T}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ceiling, $J^{V\to T}$	0.992	0.993	0.993	0.994	0.994	0.994	0.996	0.996	0.997	0.998	0.999	0.998	0.999	0.999

Design. We compare $J_\ell^{V\to T}$ with $J_\ell^{T\to T}$ at every fitted layer, by linear CKA and by the principal angles between their leading singular subspaces. CKA is computed on the J-lens vectors, that is, the rows of $W_c\,\mathrm{diag}(g)\,J_\ell^{m\to T}$, with W_c the vocabulary-centered unembedding. It equals 1 when one set of lens vectors is an orthogonal transformation of the other up to scale. It compares the two operators as the lens reads them, without reference to any particular set of activations.

A bare similarity number is not evidence. With finite sampling, two independent estimates of the same operator do not have $\mathrm{CKA} = 1$. The only interpretable comparison is against that ceiling. We therefore accumulate two independent estimates in a single pass, splitting prompts by parity, and take the ceiling at a layer to be the CKA between the two halves' estimates of the same operator. We also calibrate against a depth yardstick, $\mathrm{CKA}(J_\ell^{T\to T}, J_{\ell+k}^{T\to T})$, so that a cross-modal difference can be expressed in units of network depth.

Result: the two operators are different, by a margin sampling noise cannot explain. On the 240-prompt fit of Qwen2.5-VL-7B, the split-half ceiling is 0.987–1.000 across all 28 layers, so at this sample size the operators are estimated almost exactly. Cross-modal CKA in the middle band averages 0.478, leaving a margin of roughly 0.5 between the observed similarity and what sampling noise alone would produce (Figure 9a, Table 7).

Against the depth yardstick, walking k layers forward along the text operator gives mean CKA 0.991 (k=1), 0.986 (k=2), 0.966 (k=4), 0.851 (k=8), and 0.751 (k=12). Even k=12 stays well above the cross-modal 0.478, so the difference between the two operators exceeds what twelve layers of depth produce within a single modality (Figure 9b). Principal angles agree: at rank 32, the mean principal cosine is 0.16–0.22 on the read-in subspace and 0.31–0.38 on the write-out subspace, so the two subspaces barely overlap.

The 32B fit does not have the sample size to answer this. At 60 prompts, the split-half ceiling for $J^{V\to T}$ collapses to 0.008–0.054 through the early and middle layers, far below the cross-modal

Table 8: Broadcast metrics, medians over nine sampled layers of Qwen2.5-VL-7B-Instruct; 2048 directions per population, all 28 heads per layer. MLP gain is reported both with and without the block's post-attention layernorm, because the source is ambiguous on this point and the choice moves a unit direction's norm by $\sqrt{d}$; the two agree. Label preservation is the median over all 28 heads.

Population	**MLP gain** (norm / raw)	**OV gain** (top 10% heads)	**Label preservation**
$J^{T\to T}$ lens vectors	1.13 / 1.18	**1.52**	$+0.0013$
$J^{V\to T}$ lens vectors	0.87 / 1.26	**1.55**	$+0.0012$
J_{rot} (spectrum-preserving)	1.06 / 1.00	1.04	-0.0046
MLP neuron output directions	1.39 / 1.44	1.07	-0.0010
Isotropic random	1.00 / 1.00	1.00	$+0.0000$

CKA of 0.327. The visual operator is simply not estimated at that sample size. We report this cell as undetermined, not as a replication or a failure to replicate. A larger sample does not settle it within our budget. Lens rows cost a small fraction of a full operator, and at 1000 prompts the median split-half cosine of the $J^{V\to T}$ rows still stays below the protocol's threshold of 0.95 at every layer from L0 to L26, weakest at 0.770 (Appendix I.3), so the visual operator of this model is not well estimated through the first two fifths of the network even at that sample size.

F.4 Is visual MJ-space content broadcast like textual content?

We ask whether the visual subspace is distributed along the token axis. Section 3.2 claims it is. This section measures it.

Design. The original work argues that the J-space is a broadcast hub: the weights are structurally oriented to distribute its contents along both of the transformer's propagation axes. We measure both, which requires no backward passes and no interpretation of readout semantics, and add the modality dimension.

Along depth, the MLP gain of a unit direction v at layer ℓ is the output norm of the layer-$(\ell{+}1)$ MLP block applied to v, normalized by the median over isotropic random directions. Along the token axis, for each attention head we measure OV gain, the mean norm of $W_{OV}v$ over directions in the population, normalized the same way. We also measure label preservation, the mean reciprocal rank of $\cos(W_{OV}v_i, v_i)$ among $\{\cos(W_{OV}v_i, v_j)\}_j$, minus the same statistic on random directions, so a head that copies everything indiscriminately scores zero. The essential control is J_{rot}: the same lens vectors under a fixed random orthogonal rotation, which exactly preserves their spectrum and pairwise geometry.

Result: MLP gain does not replicate, and the sign is reversed. The original work reports that J-lens vectors are amplified roughly $10\times$ through the workspace band, while neuron output directions stay near 1. We find J-lens vectors at 0.87–1.26 and neuron directions highest, at 1.39–1.44 (Table 8). Because the source does not specify whether the block's layernorm is included, a choice that rescales a unit direction by $\sqrt{d} \approx 60$, we computed both and they agree. We therefore report this as a genuine non-replication on this model. The most likely remaining cause is scale: 28 layers here against roughly 100 in the original.

Result: OV gain replicates, and is modality-symmetric. A small set of heads amplifies J-lens directions to 1.52–$1.55\times$ random, while J_{rot} sits at 1.04. Since the rotation preserves the spectrum and all pairwise angles, this is not a norm or conditioning artifact. Crucially, $J^{V\to T}$ (1.55) and $J^{T\to T}$ (1.52) are indistinguishable: visual lens directions get the same amplification as textual ones.

Result: label preservation does not separate the modalities. $J^{T\to T}$ scores $+0.0013$, $J^{V\to T}$ scores $+0.0012$. The two are indistinguishable, and both sit near the floor. Only the rotated control is clearly below (-0.0046).

Taken together, the token-axis results say that visual lens directions are amplified by the same heads and to the same degree as textual ones, and that this measurement does not separate the modalities in either direction (Table 9).

Table 9: Broadcast metrics at each sampled layer (L0, L3, …, L24) of Qwen2.5-VL-7B-Instruct, for the lens vectors of $J^{T\to T}$ (T) and $J^{V\to T}$ (V) and the spectrum-preserving rotation (rot); 2048 directions per population, all 28 heads per layer. Table 8 gives the medians over these layers.

	OV gain (top 10%)			Label preservation			MLP gain, norm			MLP gain, raw		
Layer	T	V	rot	T	V	rot	T	V	rot	T	V	rot
L0	3.63	1.54	1.10	+0.0013	+0.0164	−0.0020	1.68	1.40	1.44	1.31	1.27	1.00
L3	1.67	1.18	1.03	+0.0617	+0.0182	+0.0004	0.45	0.63	1.17	1.24	1.10	1.01
L6	1.45	1.53	1.02	+0.0582	+0.0123	+0.0004	1.11	0.81	1.15	1.06	1.16	1.00
L9	2.01	1.55	1.09	−0.1447	−0.0875	−0.1227	1.02	0.56	1.00	1.08	1.12	1.00
L12	1.35	1.57	1.06	0.0000	+0.0012	−0.0254	1.13	0.96	1.03	1.18	1.35	0.98
L15	1.32	1.37	1.03	+0.0101	+0.0101	+0.0006	1.08	0.87	0.98	1.15	1.17	0.99
L18	1.62	1.85	1.04	−0.1185	−0.0139	−0.0792	1.87	1.13	1.09	1.44	1.26	1.01
L21	1.52	2.71	1.04	0.0000	0.0000	−0.0075	1.32	1.02	1.06	1.15	1.35	1.01
L24	1.20	3.21	1.03	+0.0060	−0.0025	−0.0046	3.62	0.67	1.03	2.53	1.30	1.00

Reconciling a disagreement in the literature. Cross-modal *task vectors* are reported to be modality-invariant (Luo et al., 2025), while SAE features for the same concept are reported to occupy *different directions* per modality (Lee et al., 2026). The bottleneck predicted by the workspace account accommodates both:

> Content *inside* the workspace has been verbalized and is therefore modality-invariant; content *outside* the workspace retains modality-specific geometry.

This is directly testable: stratify directions by MJ-space alignment (J-lens kurtosis κ) and measure cross-modal consistency per stratum. **Prediction: consistency increases monotonically with κ.** A flat relationship falsifies the reconciliation.

Result: the prediction holds, and survives the control that matters. For each vocabulary entry i we compare its J-lens vector under the two operators, $\cos\big((W_U\mathrm{diag}(g)J^{V\to T})_i, (W_U\mathrm{diag}(g)J^{T\to T})_i\big)$. Vocabulary directions are far from isotropic, so a bare cosine is uninterpretable; we subtract the same statistic under a random row permutation, which is the value expected with no correspondence at all. Stratifying by κ into quintiles over 20,000 sampled entries per layer, this excess consistency rises from 0.106 to 0.148 (mean over L9–L17, lowest to highest quintile), monotonically in four of five layers.

Two controls decide whether that is about MJ-space alignment. A random stratifier gives a span of 0.0016—flat, as it must be if the machinery is sound. A *norm* stratifier, however, gives a span of 0.0407, as large as κ's 0.0395 but running in the opposite direction, which raises the possibility that κ is only proxying for vector norm. It is not: the two are nearly uncorrelated ($\rho = -0.076$), and stratifying by κ *within* norm quintiles leaves its effect essentially unchanged at +0.0372. The two stratifiers pick up different structure, and the reconciliation's prediction survives the one that could have explained it away.

We report this as support for the direction of the reconciliation, not as a settled result: the effect is small in absolute terms, and κ is a proxy for workspace membership rather than a measurement of it.

F.5 Response of the subspace to a visual morph

We ask whether the subspace responds to a perceptual change as a step. Section 3.2 claims it does. This section measures it.

Design. We port the ambiguity experiment of the original work to vision. Where they mixed token embeddings as $(1-\alpha)e_B + \alpha e_A$, we sweep α over a pixel-space blend between two images of different categories (`cat` ↔ `dog`, `car` ↔ `bus`, `pizza` ↔ `cake`, `horse` ↔ `cow`; 11 values of α, 6–10 image pairs per category pair). At each layer we measure two projection shares of the visual-position mean activation: one along the pure-B →pure-A line, independent of the lens, and one restricted to the 2D subspace spanned by the two concepts' J-lens vectors.

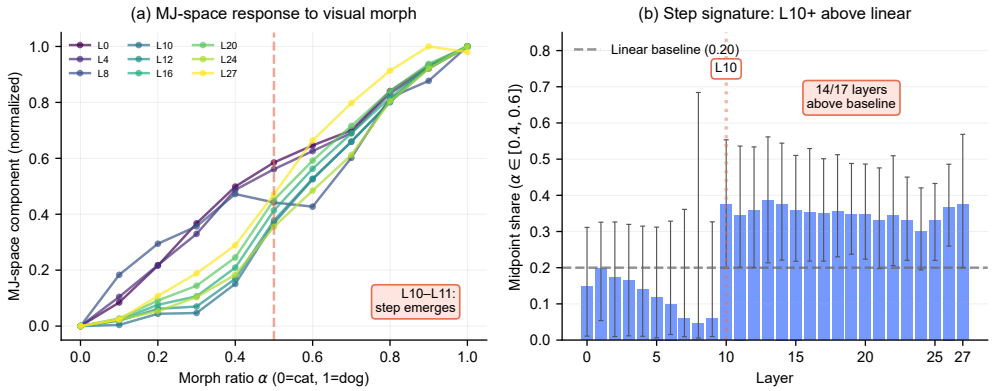

Figure 10: **From L10 the MJ-space component responds to a visual morph as a step, while the full residual stream responds linearly at every depth.** (a) MJ-space component against the morph ratio α between two category images, by layer. Below L10 the response is sub-linear—the two concepts' lens coordinates barely move across the middle of the morph—and above it the curve squares off. (b) The same data as one number per layer: the share of the curve's total change falling in the ambiguous middle ($\alpha \in [0.4, 0.6]$), against the linear baseline of 0.20 (dashed). The MJ-space component is at 0.05–0.20 through L0–L9 and steps to 0.30–0.39 from L10, exceeding the baseline with a bootstrap interval excluding 0.20 in 14 of the 17 layers above L10. The full residual stream instead sits at 0.24–0.26 at every layer including L0, so the step belongs to the MJ-space component and not to the activation as a whole. L10 is also where effective dimensionality departs baseline at text positions, three layers after it does at visual positions (Appendix F.2).

Table 10: Response to a visual morph, Qwen2.5-VL-7B-Instruct. Entries are the middle-fifth change fraction with a bootstrap 95% CI over image pairs; a linear response gives 0.20. Layers grouped for compactness; the transition layer L10 is shown separately.

Layers	**Full residual stream**	**MJ-space component**	> 0.20**?**
L0–L9	0.24–0.25	**0.05–0.20**	no
L10	0.26	0.37 [0.20, 0.55]	borderline
L11–L27	0.25–0.26	**0.30–0.39**	yes, $14/17$ layers

Statistic. We report the fraction of the curve's total change that falls in the middle fifth of the α range, $\alpha \in [0.4, 0.6]$, after normalizing each curve to $[0, 1]$. A perfectly linear response gives exactly 0.20; a step response approaches 1.

Why the visual version is stronger. Perceptual ambiguity is a native phenomenon with a century of psychophysics behind it (bistable perception, binocular rivalry), whereas embedding mixtures are artificial.

Result: the MJ-space component changes abruptly at L10; the residual stream does not. The full residual stream sits at 0.24–0.26 at every layer including L0, a small, uniform supra-linearity with no depth structure. This is expected, because a pixel-space blend is not linear in the vision encoder's feature space. The MJ-space component instead is sub-linear through L0–L9 (0.05–0.20) and then steps up to 0.30–0.39 from L10 onward, exceeding the linear baseline with a bootstrap interval excluding 0.20 in 14 of the 17 layers above L10 (Figure 10, Table 10). The step is a factor of six at a single layer boundary.

L10 is close to where effective dimensionality departs its early baseline, at L7 for visual and L10 for text positions (Table 5). Two measurements agree on the band.

Control: the transition is not a subspace artifact. A natural objection is that the 2D MJ-space subspace could be degenerate in early layers: if v_A and v_B were nearly parallel there, the projection would be noise and would look sub-linear for trivial reasons. It is not. The mean $|\cos(v_A, v_B)|$ is 0.62 at L0 and 0.59 at L10, essentially flat across the transition, and the fraction of the A–B activation difference captured by the subspace rises only gradually, from 0.119 at L0 to 0.157 at

Table 11: Occupancy K^* (median over positions) against both controls, and the share of residual-stream variance the true dictionary explains at $K = 25$, at the layers on which occupancy was estimated. At most 1% of positions is censored at the ceiling $K = 128$ in any cell.

Qwen2.5-VL-7B-Instruct

Layer	L0	L4	L8	L12	L16	L20	L24
K^* text, rotated control	39	9	12.5	5	2	9	1
K^* visual, rotated control	18	18	21	18	26.5	16.5	54
K^* text, activation-aligned	39	12	10	10	12	13	17
K^* visual, activation-aligned	19	22	21	17	24	43	55
variance explained at K=25, text	0.053	0.021	0.012	0.010	0.006	0.008	0.017
variance explained at K=25, visual	0.033	0.035	0.033	0.029	0.018	0.025	0.033

Qwen3-VL-32B, every eighth layer

Layer	L0	L8	L16	L24	L32	L40	L48
K^* text, rotated control	21	18	11	9	10	7	7
K^* visual, rotated control	25	15	7.5	18	11	9	3

L10. A 30% smooth change in subspace geometry cannot produce a sixfold step in the response shape.

The same diagnostic gives an independent echo of the F.2 result: the captured fraction collapses to 0.085, 0.074, 0.056, 0.028 at L24–L27, meaning the visual concept subspace stops carrying the distinction near the output.

Limitations. Four category pairs and 6–10 image pairs each. The effect size, while a clean step, is modest in absolute terms (0.35 against a linear 0.20). We did not measure per-trial bimodality successfully, so the all-or-none claim rests on the shape of the mean curve, not on a bimodal distribution.

F.6 Occupancy and capacity

We ask how many directions the subspace uses and how much variance it carries. Section 3.2 uses both quantities. This section defines them and reports them.

Occupancy: what the criterion does and does not resolve. The original work defines occupancy K^* as the sparsity at which the marginal reconstruction gain of the true dictionary falls below that of a same-size control of random directions, and reports a plateau near 25 across the workspace band. Two implementation points matter.

First, "marginal gain" has two readings and they disagree. The absolute reading, $\Delta(k) = \mathrm{FVE}(k) - \mathrm{FVE}(k-1)$, is systematically biased downward for a good dictionary: once the true dictionary has explained most of the variance, its absolute increments must approach zero while the control is still climbing, so the criterion fires early. On a synthetic problem with three planted atoms, the absolute reading returns $K^* \in \{3, 2\}$ and the relative reading $\rho(k) = 1 - \|r_k\|^2/\|r_{k-1}\|^2$ returns $\{3, 5\}$. Only the relative reading can produce the reported plateau, so we adopt it.

Second, the criterion is uninformative in early layers if the tap is wrong. At $\ell = 0$, the tap is the embedding output, J_0 is close to identity, and the dictionary $W_U \,\mathrm{diag}(g)\, J_0$ is therefore close to the unembedding itself, which reconstructs embedding-like activations far better than any random control at every K. With the correct tap, censoring drops to 0% at every layer and both modalities, and the estimator resolves throughout.

Against the rotated control, occupancy is higher at visual than at text positions at every layer except $\ell = 0$, and the gap is largest at L24: 18 vs. 9 at L4, 21 vs. 12.5 at L8, 18 vs. 5 at L12, 26.5 vs. 2 at L16, 16.5 vs. 9 at L20 and 54 vs. 1 at L24 (Figure 8b, Table 11). The text figure declining toward 1 is consistent with the terminal motor convergence discussed in Appendix F.2: near the output, the text-position MJ-space collapses onto essentially a single direction, the model's own next-token prediction. That decline depends on the control: against the activation-aligned control of Table 11, which preserves each direction's projected activation energy, visual still exceeds text at every layer but L0, while the text side stays between 10 and 17 from L4 on and does not decline. In either case our profile is not a replication of the reported rise-to-plateau, and we do not present it as one.

Capacity, measured without a control. The share of activation variance carried by the MJ-space needs no control to state. At $K = 25$, the true dictionary explains $0.053 \rightarrow 0.006$ of residual-stream variance at text positions and $0.033 \rightarrow 0.018$ at visual positions across L0–L16, decreasing with depth on both sides but far faster on the text side. This reproduces the qualitative capacity claim of the original work—that model activations are dominated by information outside the J-space—directly, and from L4 onward the visual share exceeds the textual one, by 0.011–0.021 (Table 11).

G Does nameability gate the workspace?

This appendix supplements §3.3. It gives the designs, tables, and controls for the ablation experiments summarized there. Here, *gated* means that only content whose discriminative property satisfies the nameability criteria in Appendix C.2 can depend on the workspace.

G.1 Whether the identified content has a causal role

Before testing whether nameability determines which tasks depend on the workspace, we need to know that the sparse decomposition in Appendix C.1 recovers directions the model actually uses. If it only recovered directions that correlate with the image, the later ablation tests would not be informative. We test this on the captioning corpus, without adding new stimuli.

Design. At visual positions in the layer band L9–L17, we project the residual stream onto the orthogonal complement of the span of the $K = 25$ lens vectors that sparse non-negative pursuit selects from $J^{V \rightarrow T}$, and let the model caption the image. The control removes K directions drawn at random from the same dictionary, not K isotropic random directions: the dictionary rows are strongly non-orthogonal and have an unusual norm distribution, so an isotropic control would be too weak and would inflate the effect. The support is taken from the unintervened forward pass at every layer, so that the removed set does not depend on earlier interventions, and differences are paired by image. Two further conditions apply the same ablation elsewhere. The text condition ablates the 25 text positions, with $J^{T \rightarrow T}$. The count-matched condition ablates a random subset of visual positions, drawn with a fixed seed and matched in count to the text positions, because a captioning prompt has 216–256 visual positions (mean 239) and a difference between the first two conditions could otherwise come from intervening at more positions. We report content-word F_1 against the human captions, and unigram overlap with the unintervened generation, which measures how much the intervention changed what the model said, independent of whether the result is still correct.

Corpus, band and dictionary size. We use the first 20 images of the fitting corpus in Appendix D, COCO val2017 images with their human captions, and the operators of the 240-prompt fit. The prompt is *Describe this image.*, and the model generates 32 tokens greedily in every condition. Content words are lower-cased alphabetic words of at least three letters outside a fixed stop list, and F_1 is computed against the union of the image's human captions. The band L9–L17 is the middle third of the 28 fitted layers, the default of our analysis code, fixed before this experiment and not tuned on it. $K = 25$ is the occupancy plateau reported by Gurnee et al. (2026), so the removed subspace is comparable in size with theirs; it is also the size used by the other ablations in this appendix.

Result: the identified directions have a causal role. Removing the identified subspace at visual positions changes the generated caption substantially more than removing an equally large subspace drawn from the same dictionary. The paired difference in overlap with the clean caption is -0.224, with a CI excluding zero (Table 12, Figure 3a). This upgrades the naming result of Appendix F.2 from correlational to causal.

Result: the effect is specific to visual positions. Text positions show no such effect ($+0.059$, CI crossing zero). The count-matched condition keeps the effect at -0.108 (CI $[-0.208, -0.007]$). This is roughly half the full-visual effect and is still separated from the text condition. The visual specificity is therefore a property of the pathway, not of how many positions were touched. We note that the count-matched interval is close to zero, which makes this the weakest link in the argument.

Table 12: Ablation of the lens-selected subspace on 20 images, band L9–L17, $K = 25$; bootstrap 95% CIs over images. *MJ-space* removes the pursuit-selected subspace, *random* removes K directions drawn from the same dictionary. The paired differences are the values of Figure 3a.

Condition	**Content-word F_1**	**Overlap with clean**
clean	0.246 [0.199, 0.296]	1.000
MJ-space, visual pos.	0.168 [0.130, 0.207]	**0.379 [0.334, 0.424]**
random, visual pos.	0.225 [0.184, 0.269]	**0.603 [0.504, 0.715]**
MJ-space, text pos.	0.229 [0.180, 0.277]	0.703 [0.598, 0.804]
random, text pos.	0.246 [0.198, 0.300]	0.644 [0.534, 0.754]
MJ-space, visual (count-matched)	0.240 [0.195, 0.286]	0.748 [0.641, 0.850]
random, visual (count-matched)	0.252 [0.204, 0.303]	0.856 [0.750, 0.947]
Paired difference (MJ-space − random)		
visual positions (all, mean 239)	−0.057 [−0.086, −0.026]	**−0.224** [−0.339, −0.105]
text positions (25)	−0.017 [−0.049, +0.007]	+0.059 [−0.053, +0.166]
visual, count-matched (25)	−0.012 [−0.028, +0.005]	**−0.108** [−0.208, −0.007]

One corpus limitation bounds the text comparison. Of the 25 text positions in a captioning prompt, only the four tokens of *Describe this image.* are not chat-template tokens, such as <|im_start|>, `system`, and `You are a helpful assistant`. The text condition therefore ablates positions that carry almost no content. The stimulus-controlled experiments below intervene at visual positions only and do not rely on this comparison.

Result: accuracy falls at visual positions. The ablation has a real accuracy cost at visual positions (content-word F_1 -0.057, CI excluding zero). This cost is absent at text positions and absent under the count-matched control. The caption both changes and becomes less accurate.

G.2 Nameability and reasoning depth do not organize ablation selectivity

The verbalizability hypothesis predicts that a task depends on the workspace when its discriminative property can be named, so ablating the workspace should impair nameable tasks more than unnameable ones. A second account predicts that tasks requiring relational or multi-step reasoning depend on the workspace more than tasks requiring a single low-level judgment. Neither prediction is supported. Across five synthetic stimulus classes, nameable and unnameable discriminations do not differ in how much more the MJ-space ablation harms them than a random control does: the difference is $+0.009$ against a permutation null with standard deviation 0.036, $p = 0.81$. Across nine visual tasks, the class-level test of reasoning depth fails in the direction opposite to the prediction, $p = 0.97$. Ablation selectivity is present and survives matched controls, but it is carried by particular stimulus classes and tasks rather than by either predicted grouping.

Shared ablation and readout. Both experiments use the intervention of Appendix G.1. At every visual position and every layer of L9–L17, we project the residual stream off the span of the lens vectors that the pursuit selects from $J^{V\to T}$ on the unintervened forward pass, $K = 25$ of them in the nameability contrast and $k \in \{5, 10, 25\}$ in the battery. The control removes the same number of lens vectors drawn at random from the same dictionary. Selectivity is $\mathrm{AUC}_{\text{random}} - \mathrm{AUC}_{\text{MJ}}$, positive when MJ-space ablation hurts more than the control. Every task asks for a one-word answer, and we score the continuous margin between the two answer tokens at the answer position, summarized by AUC over trials rather than by accuracy. On the same-or-different stimuli the model answers *same* to most trials, so accuracy is fixed at 0.5 by the balanced design (Appendix D).

Nameability contrast: stimuli and prompt. Each of the five classes is a same-or-different discrimination between two 192×192 panels shown side by side on a light background, with 32 trials per cell, half *same* and half *different*. The nameable classes are colour (red against blue) and arrow direction (left against right). The unnameable classes are random-polygon configuration, Gabor orientation, and aspect ratio. Appendix C.2 gives the nameability criteria and the class assignment. In the nameable classes, the two values are always at full range: a red or a blue square of the same saturation, and a strictly horizontal arrow pointing left or right. Difficulty is set by visibility alone, through pixel noise, blur, and shape size, so making a trial harder never makes its values harder to name. In the unnameable classes, difficulty is set by the size of the difference: the perturbation

Table 13: Every cell of the nameability contrast: clean AUC, AUC under MJ-space and under random ablation, and selectivity, 32 trials per cell. The two dropped cells have clean AUC below the floor of 0.60 (Gabor level 1 at 0.5996) and are excluded from Table 14 and Figure 3.

Class	Level	clean	MJ-space	random	selectivity	
Arrow	1	0.770	0.615	0.688	+0.072	
	2	0.932	0.730	0.801	+0.070	
	3	0.771	0.604	0.693	+0.090	
	4	0.758	0.697	0.766	+0.068	
Colour	1	0.930	0.977	0.881	−0.096	
	2	0.973	0.953	0.906	−0.047	
	3	0.809	0.932	0.791	−0.141	
	4	0.760	0.645	0.652	+0.008	
Gabor	1	0.600	0.535	0.549	+0.014	*dropped*
	2	0.775	0.744	0.764	+0.020	
	3	0.971	0.896	0.957	+0.061	
	4	0.795	0.734	0.828	+0.094	
Aspect ratio	1	0.707	0.812	0.727	−0.086	
	2	0.781	0.846	0.807	−0.039	
	3	0.648	0.816	0.709	−0.107	
	4	0.762	0.811	0.844	+0.033	
Polygon	1	0.637	0.646	0.592	−0.055	
	2	0.609	0.562	0.521	−0.041	
	3	0.539	0.441	0.471	+0.029	*dropped*
	4	0.605	0.537	0.600	+0.062	

Table 14: Selectivity of the nameability contrast on the *calibrated* stimulus set, Qwen2.5-VL-7B-Instruct, 32 trials per cell, ablation band L9–L17, K=25, scored by AUC on the continuous margin. Difficulty levels were chosen by a prior clean-only sweep targeting clean AUC in $[0.70, 0.90]$ (Appendix D); the realised range was 0.539–0.973 and cells falling below 0.60 were dropped as at floor, which is why two classes contribute three difficulty levels rather than four. Selectivity is $\mathrm{AUC}_{\mathrm{random}} - \mathrm{AUC}_{\mathrm{MJ}}$, positive when MJ-space ablation hurts more than a size-matched random ablation. Bootstrap CIs are over difficulty levels within a stimulus class.

Stimulus class	**Nameable?**	**Mean selectivity**	**95% CI**	**Sign by level**
Arrow (left/right)	yes	$\mathbf{+0.075}$	$[+0.069, +0.085]$	++++
Colour (red/blue)	yes	$\mathbf{-0.069}$	$[-0.118, -0.018]$	−−−+
Gabor orientation	no	$\mathbf{+0.058}$	$[+0.020, +0.094]$	+++
Aspect ratio	no	-0.050	$[-0.097, +0.003]$	−−−+
Random polygon	no	-0.011	$[-0.055, +0.063]$	−−+
Mean, nameable		$+0.003$		
Mean, unnameable		-0.006		
Difference		$+0.009$	permutation $p = 0.81$	

of a seven-vertex polygon's radii, the angle between two Gabor patches, and the relative difference between two ovals' width-to-height ratios. Each class has four difficulty levels, chosen by a prior sweep on clean images to put clean discrimination between AUC 0.70 and 0.90. Levels that fall below 0.60 are dropped as at floor. Every class uses the same question, so the classes differ only in the image:

> *Look at the two panels. Are the shapes in them the same or different? Answer with exactly one word: same or different.*

The margin is $\mathrm{logit}(\text{“ same”}) - \mathrm{logit}(\text{“ different”})$. Table 13 gives every cell.

Result: the nameability contrast is null. The prediction was that nameable discriminations lose accuracy under ablation at every difficulty level while unnameable ones are approximately unaffected, the multimodal analogue of the dissociation Gurnee et al. (2026) report: information that influences behavior without passing through the workspace. Every retained cell is in a measurable range. The nameable conditions give mean selectivity $+0.003$ and the unnameable conditions -0.006. The difference of $+0.009$ is a quarter of the permutation null's standard deviation (0.036), and a label-permutation test gives $p = 0.81$ (Table 14, Figure 3b). This is a real null. The instrument has the range to detect an effect, so the null is not a measurement failure.

Result: the selectivity that is present tracks the stimulus class. The aggregate null conceals large and consistent per-class effects. Arrow discrimination shows positive selectivity at all four difficulty levels ($+0.075$, CI excluding zero): removing the identified MJ-space directions costs more than removing random directions from the same dictionary. Gabor orientation does the same ($+0.058$, CI excluding zero), and Gabor is one of our unnameable conditions. Colour, which is maximally nameable, shows a significant negative selectivity (-0.069). The two nameable classes have opposite signs. Nameability therefore does not organize the data. Stimulus identity does (Figure 3c). We note a post-hoc pattern and label it as such. The two classes with positive selectivity, arrow and Gabor, are both orientation judgments, and colour, aspect ratio, and polygon shape are not. This experiment cannot test that pattern. It was not designed to, and with five stimulus classes the grouping could easily be chance.

Reasoning-depth battery: tasks and classes. The battery has nine tasks. Each asks for a one-word *yes* or *no*, scored by the margin $\mathrm{logit}(\text{“ yes”}) - \mathrm{logit}(\text{“ no”})$, and ground truth is exactly balanced by construction ($i \bmod 2$) rather than sampled. Table 15 gives the question for every task. The class assignment is fixed in code before any measurement, which is what licenses the class-level test. The predicted automatic class contains single-step, low-level tasks that do not require holding information across steps: colour naming, brightness comparison between two panels, presence of a rendered word (OCR), and image–caption match on COCO images. The predicted flexible class contains relational, multi-step tasks or tasks that require comparison across positions: a left-of relation between two objects that must be bound to their colours and shapes, a multi-hop relation over five objects (“is the shape immediately right of the red square a blue circle?”), a labelled bar-chart comparison (locate two labels, then compare), and numerosity comparison between two dot fields. Every relational task keeps the queried objects present and varies only their arrangement. Without this control, a question such as “is there a blue circle at all?” would solve the multi-hop task, and the task would silently degenerate into single-feature detection. The ninth task is an externalized arm of the multi-hop task, whose prompt lists the scene in colour-alphabetical rather than spatial order. It is the multimodal analogue of the finding of Gurnee et al. (2026) that GSM8K solved with explicit chain-of-thought is markedly more ablation-robust than the same problems answered directly: encoding the image into propositions is externalized, while sorting and composing the relation is not. A left-to-right listing would make the answer directly readable.

Reasoning-depth battery: difficulty, doses and controls. Difficulty is calibrated, not assumed. Every task was calibrated by a clean-only sweep targeting clean AUC in $[0.70, 0.90]$, as for the stimulus classes above. Each task has one visibility knob, except image–caption match, whose distractor-caption knob saturated at AUC 0.955 even with the most lexically overlapping distractor available and which therefore also degrades the image. In the two ablation runs, clean AUC is 0.65–0.95 (Tables 16 and 17). The main run has 40 trials per task. It projects out $k \in \{5, 10, 25\}$ identified directions, all taken from one $K{=}25$ pursuit whose support is ordered by selection, so $k{=}5$ and $k{=}10$ are its prefixes, and pairs each dose with a control that removes k random atoms of the same dictionary. A genuine effect should deepen monotonically with dose, while non-specific damage grows in both arms together. A separate run at $k{=}10$, with 24 trials per task, adds the rotated control: the dictionary is right-multiplied by a random orthogonal matrix, which preserves row norms and Gram spectrum exactly, and the same greedy pursuit is run on it. Dimension, norms, selection procedure and alignment with the activations are then all matched, and only the correspondence to vocabulary tokens differs.

Result: the class-level prediction of the battery fails. Against the rotated control at $k{=}10$, two tasks show positive selectivity (Table 16, Figure 11). Brightness comparison gives $+0.375$ $[+0.19, +0.59]$: ablation drives it from AUC 0.899 clean to 0.514, chance, while the rotated control leaves it at 0.889, and the main run replicates the effect at all three doses ($+0.401$, $+0.371$, $+0.319$ for $k = 5, 10, 25$). Colour naming gives $+0.111$ $[-0.03, +0.27]$, with an interval that includes zero. Every other task is flat or negative. Both positive tasks were labelled automatic in advance. The class-level test therefore fails, and it fails in the direction opposite to the prediction. In the main run at $k{=}25$, mean selectivity is $+0.095$ for the automatic class and -0.034 for the flexible class, and the exact permutation test over all $\binom{8}{4} = 70$ label partitions of the eight classified tasks gives $p = 0.97$ against the predicted ordering.

Table 15: The question of each task of the ablation battery, as given to the model after the image. Every question ends with *Answer with exactly one word: yes or no.*, omitted here. The externalized arm prefixes the multi-hop question with the scene listed in colour-alphabetical order.

Task	Class	Question
colour naming	auto	Is the square in the image red?
brightness comparison	auto	Is the left panel brighter than the right panel?
OCR presence	auto	Does the image contain the word CAT?
caption match	auto	Caption: *a COCO caption*. Does this caption describe the image?
left-of relation	flex	Is the red square to the left of the blue circle?
multi-hop relation	flex	Is the shape immediately to the right of the red square a blue circle?
chart comparison	flex	Is bar C taller than bar A?
numerosity comparison	flex	Does the left panel contain more dots than the right panel?
multi-hop, externalized	—	In this image: *one clause per object in colour-alphabetical order, e.g. the blue circle is in the second position; the red square is in the leftmost position; ...*. Is the shape immediately to the right of the red square a blue circle?

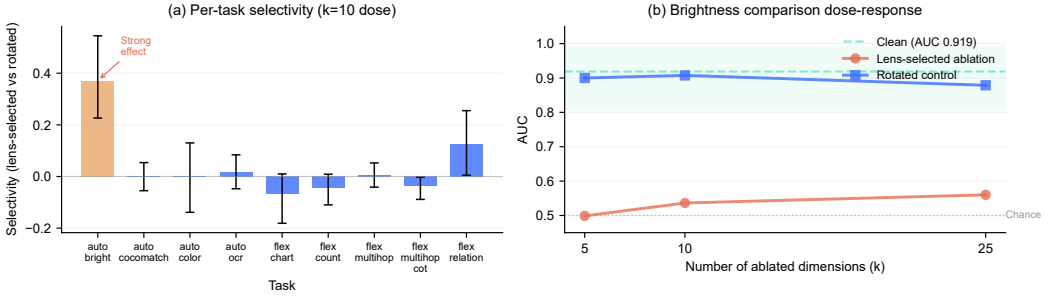

Figure 11: **Which of nine visual tasks ablation harms does not follow reasoning depth; one low-level judgment, brightness comparison, dominates.** (a) Selectivity of ablating the lens-selected subspace against the rotated control at the k=10 dose. Brightness comparison is at $+0.375$ $[+0.19, +0.59]$. Chart comparison runs significantly negative (-0.142 $[-0.32, -0.01]$), meaning ablation *improves* it. Both positive tasks had been labelled "automatic" in advance, so the class-level prediction fails, and fails in the direction opposite to the one predicted (exact permutation $p = 0.97$, main run). (b) Dose–response for brightness comparison. Ablating the identified directions drives AUC from 0.899 clean to 0.514, chance, while the control matching dimension, row norms, Gram spectrum and greedy selection leaves it at 0.889. Damage is already maximal at k=5, so the case for this effect rests on its size against a matched control and on replication across two runs rather than on monotonicity in dose.

Our class assignment was internally inconsistent on exactly the pair that matters. We assigned brightness comparison to automatic and numerosity comparison to flexible, but these are the same task structure: two panels, one magnitude, one comparative question. Our own stated criterion for the flexible class, "requiring comparison across positions," covers both. The two tasks behave completely differently ($+0.375$ vs -0.028), so task structure does not organize these data. What differs between the two is the perceptual dimension judged, which is consistent with the hypothesis stated at the end of this subsection but does not test it. Our labels were not measuring what we said they measured. We report the class test because it was fixed in advance. It should be read as a failed operationalization, not as evidence about the flexible/automatic distinction.

Removed activation energy does not explain the effect. In the rotated-control run, MJ-space ablation removes 2.2–3.7% of the visual-position activation energy, while both controls remove only 0.6–0.9%. The arms are matched in dimension but not in removed energy. Removed energy nonetheless fails to predict damage, and it predicts the wrong sign. Across the nine tasks, the rank correlation between removed energy and AUC loss is $\rho = -0.50$: the tasks whose activations lose the most energy are damaged the least. The largest removed energy is the externalized multi-hop condition (3.7%), and its AUC is unharmed ($0.788 \rightarrow 0.830$). Brightness comparison loses less energy (2.6%) and collapses to chance. What distinguishes the arms is not how much was taken out but which directions were taken out.

Ablation makes several tasks better, not worse. Chart comparison rises from 0.833 clean to 1.000 under MJ-space ablation, and its selectivity against the rotated control is significantly negative

Table 16: Ablation battery, Qwen2.5-VL-7B-Instruct, $k = 10$ directions removed at visual positions in layers L9–L17, 24 trials per task. All entries are AUC. The **rotated control** is the decisive one: the dictionary is right-multiplied by a random orthogonal matrix—preserving row norms and Gram spectrum exactly—and the *same* greedy pursuit is run on it, so dimension, norms, selection procedure and alignment with the activations are all matched and only the correspondence to vocabulary tokens differs. Class labels were fixed in code before measurement.

Task	Class	clean	MJ-space	random	rotated	sel. vs rotated [95% CI]	energy
brightness cmp.	auto	0.899	**0.514**	0.913	0.889	$+\mathbf{0.375}$ $[+0.19, +0.59]$	0.026
colour naming	auto	0.712	0.608	0.722	0.719	$+0.111$ $[-0.03, +0.27]$	0.022
OCR presence	auto	0.649	0.618	0.635	0.653	$+0.035$ $[-0.05, +0.13]$	0.023
caption match	auto	0.767	0.781	0.753	0.757	-0.024 $[-0.15, +0.08]$	0.035
left-of relation	flex	0.896	0.747	0.802	0.688	-0.059 $[-0.26, +0.13]$	0.024
multi-hop	flex	0.948	1.000	0.986	0.955	-0.045 $[-0.14, +0.00]$	0.033
chart cmp.	flex	0.833	1.000	0.910	0.858	-0.142 $[-0.32, -0.01]$	0.029
numerosity cmp.	flex	0.795	0.819	0.799	0.792	-0.028 $[-0.12, +0.05]$	0.024
multi-hop, scene externalized	—	0.788	0.830	0.792	0.792	-0.038 $[-0.13, +0.03]$	0.037

Table 17: Dose–response of the ablation battery in the main run: selectivity ($\mathrm{AUC}_{\mathrm{random}} - \mathrm{AUC}_{\mathrm{MJ}}$) with bootstrap 95% CIs at $k = 5$, 10 and 25 directions removed at visual positions in L9–L17, 40 trials per task, against a control that removes k random atoms of the same dictionary. The rotated-control run of Table 16 is a separate run at $k = 10$ with 24 trials. Clean AUC is that of the main run.

Task	Class	clean	$k = 5$	$k = 10$	$k = 25$
brightness cmp.	auto	0.919	$+0.401$ $[+0.25, +0.57]$	$+0.371$ $[+0.23, +0.55]$	$+0.319$ $[+0.18, +0.49]$
colour naming	auto	0.770	-0.005 $[-0.09, +0.09]$	$+0.001$ $[-0.14, +0.13]$	$+0.047$ $[-0.06, +0.16]$
OCR presence	auto	0.654	$+0.026$ $[-0.05, +0.10]$	$+0.016$ $[-0.05, +0.08]$	$+0.035$ $[-0.02, +0.09]$
caption match	auto	0.836	$+0.010$ $[-0.05, +0.07]$	$+0.000$ $[-0.05, +0.05]$	-0.020 $[-0.08, +0.03]$
left-of relation	flex	0.881	$+0.042$ $[-0.09, +0.18]$	$+0.127$ $[+0.005, +0.26]$	-0.069 $[-0.22, +0.08]$
multi-hop	flex	0.931	$+0.009$ $[-0.05, +0.07]$	$+0.005$ $[-0.04, +0.05]$	$+0.004$ $[-0.05, +0.06]$
chart cmp.	flex	0.887	-0.036 $[-0.14, +0.04]$	-0.069 $[-0.18, +0.01]$	-0.057 $[-0.16, +0.02]$
numerosity cmp.	flex	0.662	-0.031 $[-0.11, +0.03]$	-0.046 $[-0.11, +0.01]$	-0.015 $[-0.09, +0.05]$
multi-hop, externalized	—	0.859	-0.029 $[-0.09, +0.02]$	-0.037 $[-0.09, -0.002]$	-0.035 $[-0.08, -0.005]$

(-0.142 $[-0.32, -0.01]$). Multi-hop rises from 0.948 to 1.000. We do not have a tested account of this. A natural reading is that removing verbalizable content removes a competing verbal heuristic that was interfering with a comparison the model could otherwise make perceptually. This is a hypothesis, not a result.

The dose–response check does not support the small effects. Only 1 of 9 tasks has selectivity that increases monotonically with k, against 1.5 expected by chance (3 doses admit 6 orderings, of which one is increasing). We therefore decline to treat any of the sub-0.1 selectivities as effects. Left-of relation reaches $+0.127$ $[+0.005, +0.26]$ at $k = 10$ in the main run, but it changes sign across doses and is negative against the rotated control (-0.059), so we do not treat it as an effect either. Brightness comparison does not deepen with dose: the damage is already maximal at $k = 5$ and recovers slightly by $k = 25$. This is consistent with five directions being sufficient to remove the judgment, but monotonicity provides no support for it. Its case rests on effect size against a matched control and on replication across two runs. Table 17 gives every task at every dose in the main run.

Result: the externalized arm is uninformative. The externalized arm is more ablation-robust than the direct arm (-0.035 vs $+0.004$ in the main run at $k = 25$, a difference of $+0.039$ in the predicted direction). But the direct arm shows no MJ-space dependence to begin with ($+0.004$ $[-0.045, +0.062]$), so there is nothing for externalization to protect. A robustness comparison between two conditions that are both already robust cannot test the externalization account, which needs a multimodal task that is itself workspace-dependent. On the evidence of Table 16, the relational tasks we built are not.

What this does and does not rule out. It does not falsify the verbalizability hypothesis. Our operationalization of "nameable" is a single-token vocabulary criterion applied to five synthetic classes. The ablation is confined to L9–L17 at K=25 for the nameability contrast and $k \leq 25$ for the battery, and 32 trials per cell give AUC intervals of roughly ± 0.18. What the two experiments establish is narrower. In this setting and with these operationalizations, the two predictions we set out to test

Table 18: Naming-installation experiment on the *calibrated* stimulus set, in-context condition only. Qwen2.5-VL-7B-Instruct, 40 trials per cell, difficulty level 2, ablation band L9–L17, $K{=}25$, scored by AUC on the continuous margin. Sensitivity is $\mathrm{AUC}_{\mathrm{random}} - \mathrm{AUC}_{\mathrm{MJ}}$; the prediction is that naming raises it. The *irrelevant* row names a different feature dimension at matched prompt length. Polygon's clean AUC is below the 0.60 floor, so its rows are shown but excluded from the means.

Stimulus	**Prompt**	**AUC clean**	**AUC MJ**	**Sens.**	**Δsens vs. none**
Aspect ratio	none	0.685	0.741	−0.054	—
	named	0.869	0.873	−0.026	+0.027
	irrelevant	0.614	0.691	−0.026	+0.027
Gabor	none	0.889	0.791	+0.007	—
	named	0.940	0.906	+0.010	+0.003
	irrelevant	0.922	0.835	+0.030	+0.023
Polygon	none	0.570	0.481	−0.010	—
	named	0.573	0.535	+0.025	+0.035
	irrelevant	0.589	0.464	−0.005	+0.005
Mean Δsens, *named* (aspect, Gabor)					+0.015
Mean Δsens, *irrelevant* (aspect, Gabor)					+0.025

do not hold: workspace dependence tracks neither nameability nor reasoning depth. In both experiments, substantial and control-surviving selectivity exists, but it is organized along an axis we did not fix in advance. The tasks that show it are judgments of a single low-level perceptual dimension (arrow and Gabor orientation in the nameability contrast; brightness and colour in the battery) rather than tasks requiring relational composition. The sharpest single comparison is within the nameability contrast: colour, the most nameable dimension we tested, carries selectivity −0.069, while Gabor orientation, which has no name at all, carries +0.058. If nameability were the admission criterion, those two numbers would be reversed.

We therefore state the surviving pattern as a hypothesis, not a finding: whether the workspace mediates a visual judgment may depend on which perceptual dimension the judgment is about, rather than on whether its answer has a name. We have not established it. The per-condition signs across the two experiments do not line up. Colour runs negative when the question is same-or-different and positive, with an interval including zero, when the question is naming. With five stimulus classes and nine tasks, the grouping could still be chance. A cheap decisive test is to cross perceptual dimension with question form; we have not run it.

G.3 Whether naming a feature installs it in the workspace

We take the three unnameable stimulus classes, define an arbitrary name for the discriminative feature in the prompt, and repeat the ablation of Appendix G.2 at difficulty level 2 with 40 trials per cell. The name is the nonce word *GRIBBLE*, attached to the stripe pattern's tilt angle for Gabor, the blob's corner profile for polygon, and the oval's width-to-height ratio for aspect ratio. An irrelevant-name control of the same length and form gives the name to a dimension that does not vary between the panels and keeps the ordinary matching rule. If naming installed the feature in the workspace, MJ-space ablation would begin to impair the task once the name is given. It does not. The three polygon cells have clean AUC below the floor of 0.60 and are excluded; in the others clean AUC ranges from 0.61 to 0.94, so the measurement has room to move. Averaged over aspect ratio and Gabor, naming raises the ablation sensitivity, $\mathrm{AUC}_{\mathrm{random}} - \mathrm{AUC}_{\mathrm{MJ}}$, by +0.015 and the irrelevant name raises it by +0.025. The difference, −0.010, is far below what 40 trials per cell can resolve (Table 18). This is a null for the naming-installation hypothesis as operationalized here, with the name defined in the prompt only and on one model.

H What the image channel reads out: details

This appendix supplements §3.4. It gives the validation of the visual readout, the comparison of the modality index across output contexts, and the linearization error of the visual head.

H.1 VALIDATION OF THE VISUAL READOUT

The rendered readout is computed in the image-output context of Janus-Pro-1B on the model's own rollout of 576 visual tokens for a COCO caption. At layer ℓ and each visual position i, the visual readout of Eq. (3) ranks the 16,384 codebook entries. The operator is $J_\ell^{V\to V}$, whose source modality matches the position. We keep the top-ranked entry at each position. The 576 entries are laid out on the 24×24 token grid and passed through the model's own VQ decoder. This gives one image per layer. We do the same with two other readouts of the same activations. The first is a rotated copy of the visual dictionary, which should decode to noise. The second is the model's own generation head applied directly to the intermediate activation, the visual analogue of a logit lens. Figure 4a shows the rendered readout for one caption.

We measure pixel correlation between the rendered readout and the input image. At a visual position of a rollout, the input token is the one the model emitted there. The image formed by the input tokens is therefore the model's own image. We decode that image and the image of the rendered readout. We flatten each into a vector of pixel values, standardize both, and report their Pearson correlation over all pixels. The first three captions of the corpus are rendered at six layers (Table 19). Three captions are enough to show the range. The rendered readout correlates with the input image at 0.73–0.86 at L0, before any computation has acted. The correlation falls with depth. The generation head starts near zero and ends at 0.37–0.62, not monotonically. The rotated control stays near or below zero. The readout that begins where the input is and the head that ends where the output is converge without crossing.

We use a swap control to test whether the readout at visual positions reflects the context or simply copies the tokens already present. We pair the visual tokens of one image with a different, non-matching text caption. The prompt comes from one caption, and the visual tokens come from the model's rollout for another. We decode the rendered readout at L0, L8, L14 and L22. The readout follows the visual tokens at all four layers, including L0: each renders the image the swapped-in tokens form, and none renders anything that belongs to the prompt (Figure 4b). The readout at visual positions is therefore an echo of the tokens already present, not an interpretation of the context. This control was run on one pair of captions and judged by inspecting the four decoded images.

A second measurement asks whether the top-1 codebook entry of the rendered readout matches the token the model actually emits next, rather than the token already present. Pixel correlation cannot make this distinction, because in a finished rollout the input and the output are the same image. The hidden state at position i is the one from which the model emits token $i{+}1$. A readout that predicts should therefore agree with token $i{+}1$, and a readout that copies should agree with token i. Over the 575 positions that have a successor, we report the fraction at which the top-1 entry equals token $i{+}1$, and beside it the fraction at which it equals token i (Table 19). Exact agreement is a weak statistic on this codebook, whose entries are highly redundant. Even the agreement with token i is at most 0.016 at any layer, although the decoded images are nearly the same. The rendered readout matches the next token on at most 0.004 of positions through L14 (0.000 at L0 and L8, at most 0.002 at L4 and 0.004 at L14), and on 0.017–0.035 at L22. The generation head, which does predict at the last layer, is also at 0.000 through L8. In the swap control the agreement with the next token is 0.000 at L0 and L8, 0.005 at L14 and 0.019 at L22, close to the values of the unswapped runs. We therefore use the rendered readout only qualitatively, and only at text token positions (Figure 4c). The top-20 check in Appendix H.2 measures prediction with the 20 highest-ranked entries instead of the top one, and that is where the predictive content of the visual readout shows.

At prompt positions the readout does not depict. There is no visual token there to copy. Ranking the codebook by the raw readout score returns nearly the same entries for unrelated captions, so only the differential ranking carries content. This is the same failure the raw modality index shows (Appendix C.1.1). The entries of the differential ranking decode to flat colour fields rather than object structure. That much follows from Eq. (2) and needs no measurement. The expectation in the operator runs over target positions, so $J_\ell^{T\to V}$ carries no spatial index, and any display built from it is spatially uniform whatever the model is about to draw. Spatial structure can enter only through the source side, by reading at many positions at once. That is what the rendering at visual positions does, and it is why that rendering is an echo.

Table 19: The rendered readout at visual positions, Janus-Pro-1B, image-output context, three captions; ranges are over captions. Pixel correlation is with the decoded input image (the model's own rollout). Agreement is the fraction of positions at which the top-1 codebook entry equals the token at the position (i) or the next emitted token ($i{+}1$).

	pixel correlation			rendered readout		generation head
Layer	rendered readout	generation head	rotated	agrees with i	agrees with $i{+}1$	agrees with $i{+}1$
L0	0.73–0.86	0.06–0.11	-0.26 to -0.14	0.004–0.012	0.000	0.000
L4	0.71–0.87	0.32–0.55	-0.01 to $+0.30$	0.007–0.016	0.000–0.002	0.000
L8	0.71–0.86	0.24–0.34	0.00 to $+0.02$	0.002–0.009	0.000	0.000
L14	0.69–0.78	0.19–0.51	-0.24 to $+0.20$	0.002–0.004	0.000–0.004	0.000–0.007
L18	0.60–0.77	0.40–0.67	-0.19 to $+0.09$	0.000–0.005	0.000–0.005	0.007–0.021
L22	0.52–0.76	0.37–0.61	-0.17 to -0.11	0.000–0.007	0.017–0.035	0.026–0.033



Figure 12: **The visual component of the MJ-space depends on output modality** (Janus-Pro-1B, prompt positions, same weights in both contexts). (a) Excess modality index over the rotated control, by layer and output context. The two contexts start together at L0 (0.118 image vs. 0.128 text), separate from L2 onward, and are roughly $11\times$ apart at L20 and $21\times$ at L22. Mid-band excess is 0.208 when the model is about to emit an image against 0.055 when it is about to emit text; the difference of $+0.153$ is $32\times$ the split-half noise of 0.005. (b) Why the excess statistic, and not raw MI. Raw MI runs the other way (0.611 text vs. 0.526 image), but the rotated control—which destroys every correspondence between lens vectors and codebook entries—itself returns 0.556 in the text context, 91% of the raw value. A statistic that a content-destroying control reproduces to within 9% is not measuring content.

The target positions can be kept separate. In reverse mode this costs 576 backward passes, but the cost is a property of reverse-mode differentiation, in which spreading the cotangent over all target positions is what performs the pooling. Forward mode is its dual and returns all 576 target positions from one pass. The resolution is not empty: only 3–21% of the energy of the transport is position-independent. It still does not depict the prompt (Figure 4c), and a norm-matched rotation degrades the image over the same range. This was run on a single prompt, so it establishes the cost and the form of the readout, not the generality of the null.

Colour–object binding cannot be linearly decoded from lens scores. This holds in Qwen2.5-VL-7B, which answers all 120 binding trials correctly, so the null constrains the lens, not the network.

H.2 The modality index by output context

At text token positions in the middle band, the visual component of the MJ-space is much larger when the model is about to emit an image than when it is about to emit text. The modality index (MI) measures how much of an activation's sparse decomposition lands on visual atoms. Appendix C.1 defines the index and its three controls. We compare the index across two output contexts of Janus-Pro-1B (Chen et al., 2025): one where the model is about to emit an image, and one where it is about to emit text. Averaged over L8–L14 at prompt positions, the excess index over the rotated control is 0.208 in the image-output context and 0.055 in the text-output context (Table 20). The difference, $+0.153$, is 32 times the split-half noise of 0.005. These are the band averages of §3.4. The visual readout on which the index depends passes its check: at L22 it places the model's own next emitted visual token in its top 20 codebook entries on 0.214 of positions, against a chance rate of 0.0012. Figure 12 gives the profile by layer, and Appendix C.1 gives the reason the raw index cannot be

used. The rest of this section gives the design, the check and the controls, and then what the result does and does not establish.

The two contexts share the model, the 12 tapped layers, the same corpus of 150 COCO items, the tap point, the chunking and the split-half parity. They differ in output modality and, by design, in their input. In the text-output context the input is an image, a "describe this image" instruction, and the model's own greedily decoded caption appended by teacher forcing. Its target positions are the text positions, which gives $J^{V\to T}$ and $J^{T\to T}$. In the image-output context the input is a generation prompt followed by the model's own sampled rollout of 576 visual tokens. Its target positions are the emitted visual tokens, which gives $J^{T\to V}$ and $J^{V\to V}$. The rollouts are the model's own in both contexts, because a VQ encoding of a real image would measure the Jacobian of a sequence the model would not itself produce. The prompt positions at which MI is measured are therefore the instruction and the model's caption in the text-output context, and the generation prompt in the image-output context. Only the text-output context has an image in its input, and it is the context with the smaller visual component, so the input image does not account for the difference. The two contexts also differ in the wording of their text, which this design does not control. Comparing two output modes within one model avoids the confound of scale, data and training recipe that a comparison between an understanding-only model and a unified model would carry.

We use the 1B model rather than the 7B because the effective sample size of a Jacobian estimate is the number of source positions it averages over. Our Qwen3-VL-32B replication of the operator comparison failed for this reason: its split-half ceiling (0.008–0.054) fell below the similarity being measured, so no verdict was possible (Appendix F.3). The comparison here is a difference between two estimates, so it is doubly exposed to this. At fixed compute, halving d_{model} roughly doubles the number of prompts we can afford, so we run the comparison at 1B with 150 prompts per context rather than at 7B with too few. For the same reason the generation prompt carries a fixed stylistic prefix and suffix, which lengthen it from 18 to about 40 tokens. Those tokens are both where MI is measured and the source positions of $J^{T\to V}$.

We use one readout matrix for both contexts. The visual readout is the linearized head W_{eff} of Eq. (3), defined in full in Appendix H.3. The derivative σ' of the head's nonlinearity is averaged over the positions where the visual head acts. The same W_{eff} is used in both contexts. If the readout matrix differed between arms, the difference being measured would include the change in the readout itself. Appendix H.3 measures what the linearization costs.

Because MI depends on this readout, we check the readout before we use it. At the deepest tapped layer J_ℓ is close to the identity, so a correct visual readout must predict the model's own next emitted visual token. We report its top-20 hit rate, the fraction of visual positions at which the token the model emits next is among its 20 highest-scoring codebook entries. The model's own generation head gives the upper bound and a rotated control dictionary the floor. If this check failed, no modality-index number in this section would be interpretable. The value of 0.214 at L22 has a 95% interval of $[0.197, 0.230]$ over 40 rollouts not used in any fit. The chance rate is $20/16384 = 0.0012$, the model's own generation head reaches 0.581 $[0.561, 0.600]$, and the rotated control is at 0.001. The hit rate rises with depth (0.006 at L0, 0.057 at L16, 0.214 at L22), the profile the text side shows. The first estimate of this check, on six rollouts, was 0.170 against 0.576; the forty-rollout values replace it. The check first failed at exactly zero because it applied $J^{T\to V}$ to activations at visual positions. That error and the rule it led to are recorded in Appendix D.

The three controls on the modality index are defined and reported in Appendix C.1. Table 20 gives their values in both contexts.

Positions matter. In the generation context, MI at the emitted visual token positions is high by construction, because those are the positions at which the model chooses its output. These are the motor positions. The prediction is a claim about the prompt positions in the middle band. We therefore measure both prompt positions and motor positions, and we report both.

The prediction is that in the image-output context a nonzero MI emerges across the middle band with the structure the original work reports for the J-space, and that in the text-output context MI is approximately zero.

Understanding-only models have no visual output space, so MI is undefined for them, not zero. It is tempting to complete the comparison with a Qwen-VL row at $\text{MI} = 0$. That would be a category

Table 20: The modality index by output context, Janus-Pro-1B, 150 prompts per context for the operators and 20 for the index, 12 tapped layers. MI is averaged over the middle band (L8–L14) at *prompt* positions. **Excess MI** = MI − MI(rot) is the primary statistic: the rotated control preserves row norms and Gram spectrum exactly and destroys only the correspondence to codebook entries, so anything it reproduces is dictionary geometry rather than visual content. Half-columns are the odd/even split-half estimates of the excess.

Output mode	raw MI	MI(rot)	excess MI	MI(size-matched)	$\text{half}_0/\text{half}_1$	vis. exc. kurt.
text out	0.611	0.556	0.055	0.623	0.054 / 0.056	−0.2
image out	0.526	0.318	**0.208**	0.546	0.211 / 0.206	−0.3
difference	−0.085	−0.238	**+0.153**	−0.077	noise 0.005	−0.1

error. A lens vector is a direction of first-order causal influence on an output. A model whose only output is text has no visual output space, so there is no basis in which to express a visual component of the MJ-space. MI has no value at all. For understanding-only models, the verbal format of the MJ-space is a consequence of the definition, not an empirical finding. The prediction therefore has to be tested in a model that can draw.

The component is built by the network, not present in the embeddings. At L0 the two contexts have the same excess MI: 0.118 for image output and 0.128 for text output. They separate from L2 onward and stay separated at every subsequent layer. The image-output excess climbs to 0.35 at L20, while the text-output excess falls to 0.03. This is a ratio of about 11× at L20, and of about 21× at L22, the last tapped layer (0.334 against 0.016). Whatever the visual component is, it is not a property of the input representation.

The visual component does not show all the signatures the original work reports for the J-space. The prediction has two clauses: a nonzero MI, and that structure. The first holds. The second only partly holds. The excess kurtosis of the visual readout is −0.4 to +0.2 across all layers and both contexts. It is indistinguishable from its own rotated control at every layer. On the signature that the original work uses to establish that J-lens readouts are sparse and concept-like, the visual side shows nothing. The visual readout is also an order of magnitude lower-dimensional than the text readout: at L18–L22 its effective dimensionality is 0.016–0.041 in the two contexts, against 0.29–0.64 for the text readout. Its effective dimensionality rises with depth up to L20 and changes little at L22, rather than showing the inverted-U we found at visual positions in the understanding-only model (Appendix F.2). We therefore claim only that there is a content-specific visual component whose amount tracks output modality sharply. It does not look structurally like the J-space of the original work. We report the failed signature alongside the positive one.

Motor positions are higher, as expected. Excess MI at the emitted visual token positions is 0.276, against 0.208 at prompt positions. At those positions the model is literally choosing codebook entries. This is why the claim is made at prompt positions. That the prompt positions carry three quarters as much excess as the motor positions is the substantive observation.

The limitations are as follows. We study one model at one scale. Janus-Pro-1B was chosen over the 7B specifically to buy source positions, and we have not verified that the effect survives at 7B. The visual readout linearizes a two-layer head. Appendix H.3 measures what that costs at every layer. Twenty prompts per context enter the MI estimate; the 150 prompts per context are for the operator fits. The layer-by-layer profile is therefore noisier than the band average. The excess-MI statistic follows the rotated-control subtraction used in the rest of the paper, but we selected it after we saw that raw MI was dominated by its control. A replication with the statistic fixed in advance would settle that.

H.3 Linearization error of the visual readout

The generation head of Janus-Pro maps a normalized final-layer state x to the logits of the 16,384 codebook entries:

$$\text{head}(x) \;=\; W_{\text{head}}\,\sigma(z) + b_2, \qquad z \;=\; W_{\text{proj}}\,x + b_1. \tag{9}$$

Here W_{proj} and b_1 form the first layer of the head, σ is the GELU applied to each coordinate of the pre-activation z, and W_{head} and b_2 form the second layer. The visual readout replaces the head by

the linear map $W_{\text{eff}}x$ of Eq. (3), where $W_{\text{eff}} = W_{\text{head}}\,\mathrm{diag}\big(\mathbb{E}_t[\sigma'(z_t)]\big)\,W_{\text{proj}}$. In this expression σ' is the derivative of the GELU, z_t is the pre-activation at position t, and the mean $\mathbb{E}_t$ runs over the visual positions, where the head acts, of the rollouts used to fit the operators. W_{eff} therefore keeps the average slope of the nonlinearity and drops both biases.

We measure the error of this replacement on 40 fresh Janus-Pro-1B rollouts. The rollouts use COCO val2017 captions from index 3000 onward, with 96 image positions each. None of them was used in any fit. At layer ℓ and a visual position with activation h, let $x_\ell = \mathrm{norm}(J_\ell^{V\to V}h)$ be the transported activation, where norm is the model's final normalization. We compare four readouts of x_ℓ:

$$\begin{aligned}
&\text{real head:} && \mathrm{head}(x_\ell),\\
&\text{linear:} && W_{\text{eff}}\,x_\ell,\\
&\text{affine:} && W_{\text{eff}}\,x_\ell + c, \qquad c = \mathrm{head}(\bar{x}) - W_{\text{eff}}\,\bar{x},\\
&\text{first-order:} && \mathrm{head}(\bar{x}_\ell) + W_{\text{head}}\,\mathrm{diag}\big(\sigma'(\bar{z}_\ell)\big)\,W_{\text{proj}}\,(x_\ell - \bar{x}_\ell).
\end{aligned} \tag{10}$$

In the affine readout, $\bar{x}$ is a mean input, taken either as the mean final-layer state or as the mean of x_ℓ at the same layer; the affine column of Table 21 uses the final-layer mean. In the first-order readout, $\bar{x}_\ell$ is the mean of x_ℓ and $\bar{z}_\ell = W_{\text{proj}}\bar{x}_\ell + b_1$, so the expansion re-evaluates σ' at that point. Every mean in Eq. (10) is estimated on half of the rollouts and scored on the other half. The real head applied to x_ℓ is the best a readout through this head could do with the same transport. Comparing the linear readout with it therefore isolates the linearization from the transport.

Each readout is scored by its top-20 hit rate, the fraction of positions at which the token the model emits next is among the 20 codebook entries with the highest score. The ratio in Table 21 is the hit rate of the linear readout divided by that of the real head on the same input. The overlap is the fraction of the 20 highest-scoring entries of the linear readout that are also among the 20 highest-scoring entries of the real head. The final row applies the readouts to the final-layer state itself, with no transport.

The reference for how much error is acceptable is the text J-lens itself. Its readout W_U is exactly linear, so its only approximation is the transport J_ℓ. Its next-token hit rate is the fraction of positions at which the next token of the text is among its 20 highest-scoring vocabulary entries. That rate relative to the model's own output, which reaches 0.740, measures what the approximation costs. We measure it on held-out pretraining text with the $n = 1000$ operator. Each half of the corpus is read by the operator fitted on the other half.

Table 21: The text J-lens against the linearization of the visual head, Janus-Pro-1B. Text: next-token top-20 hit rate of the J-lens and its ratio to the model's own output (0.740). Visual: next emitted token's top-20 hit rate of the real head on the transported activation, of the linear readout, and of the linear readout with the head's constant term added back (affine); the ratio of the linear readout to the real head; and the top-20 overlap between the two. The final row applies no transport. Chance is 0.0002 for text and 0.0012 for the visual codebook; the rotated controls are at most 0.002 throughout.

	text J-lens		visual readout				
layer	hit@20	of output	head$\circ J$	W_{eff}	affine	W_{eff}/head	overlap
L0	0.041	0.06	0.008	0.006	0.007	0.78	0.313
L2	0.083	0.11	0.009	0.007	0.008	0.82	0.333
L4	0.120	0.16	0.009	0.007	0.007	0.80	0.356
L6	0.117	0.16	0.011	0.007	0.008	0.63	0.368
L8	0.038	0.05	0.014	0.010	0.010	0.74	0.359
L10	0.101	0.14	0.020	0.011	0.011	0.54	0.380
L12	0.115	0.15	0.038	0.023	0.023	0.62	0.335
L14	0.181	0.25	0.056	0.033	0.034	0.59	0.328
L16	0.217	0.29	0.083	0.057	0.060	0.68	0.346
L18	0.299	0.40	0.132	0.102	0.103	0.78	0.454
L20	0.408	0.55	0.195	0.148	0.149	0.76	0.501
L22	0.506	0.68	0.271	0.214	0.212	0.79	0.555
final	0.740	1.00	0.581	0.524	0.535	0.90	0.704

The head's linearization costs less than the lens's own. At every layer the linear readout retains between 0.54 and 0.82 of the hit rate of the real head on the same input. At the final layer it retains 0.90, where no transport is involved. At the same layers the text J-lens retains between 0.05 and 0.68 of the model's own output. The approximation the visual readout adds is therefore smaller at every

depth than the one every J-lens readout already makes. The gap between the visual readout and the model's own head at L22 (0.214 against 0.581) comes mostly from the transport. The transport lowers the real head from 0.581 to 0.271. The linearization lowers it further from 0.271 to 0.214.

The error is curvature, not an omitted constant. The linear readout drops the head's constant term c of Eq. (10). This term shifts every code's logit by a fixed amount and could by itself reorder the top 20. Adding it back changes the hit rate by at most 0.011 at any layer, whether the constant is taken at the final layer or at each layer's own mean (0.524 to 0.535 at the final layer, 0.214 to 0.212 at L22). The first-order expansion recovers part of the gap at L8–L12, where every hit rate of the visual readout, the real head's included, is below 0.04. It widens the gap at L14–L22 (0.186 against 0.214 at L22). The linear readout therefore misses the curvature of the GELU across the spread of activations. No fixed matrix represents that curvature. The constant is a per-code shift that does not depend on the input, and no dictionary row carries it in any case. It is not the cause of the error.

The operating point does not matter. The fitted W_{eff} uses the mean of σ' over the positions where the head acts. Re-estimating that mean on the forty new rollouts gives a vector with cosine 0.9993 to the fitted one. Evaluating σ' at the mean pre-activation instead gives 0.9929. Neither changes the final-layer overlap with the real head (0.703 against 0.704). The modality index of §3.4, recomputed with each of them, keeps the difference between contexts at 32 to 42 times its split-half noise:

operating point	excess, image out	excess, text out	difference	split-half noise	ratio
fitted (reproduction)	0.208	0.055	+0.153	0.0048	32
$\sigma'(\mathbb{E}z)$	0.202	0.025	+0.177	0.0043	42
$\mathbb{E}\sigma'$, new rollouts	0.209	0.056	+0.153	0.0047	33

I THE WRITING PROTOCOL: DESIGN, DEVIATIONS, AND FULL RESULTS

Section 3.5 reports the protocol's results at the level of the claims they support: on Janus-Pro, Jacobian-lens rows are not suitable as write directions, since difference-of-means directions can install a concept and lens rows cannot, which shows that a readout lens does not automatically provide control directions. This appendix gives the design in full, the two points at which the runs departed from the written protocol, the calibration of every lens the protocol uses, and every cell of every experiment. The protocol was committed to the repository before the first image was generated, and each deviation was recorded before the results it affects were analysed.

I.1 DESIGN

Models and sampling. Janus-Pro-1B (24 blocks) and Janus-Pro-7B (30 blocks) generate with their native settings: classifier-free guidance at scale 5, temperature 1, 576 image tokens, no top-k or top-p truncation. Every run carries its own random generator seeded with the run's seed. The unintervened image and every intervened image of the same seed therefore share their random draws, and arms are compared on matched runs.

Concepts and prompts. Identity uses 24 concepts in six categories (animals, food, vehicles, household objects, nature, buildings) and 24 host–target pairs. The first two members of each category are paired with each other, and the third and fourth are paired with the corresponding member of the next category, so half the pairs cross a category boundary. Each pair is run in three templates: the stylised generation prompt of the earlier protocol, "A photo of a {w}.", and "A close-up picture of a {w}, centred, on a plain background." Seven property families of simple drawn or photographed objects carry the appearance side: colour (four values, on a car and on a cup), size (small or large ball), shape (circle, triangle, square, star), direction (an arrow pointing up, down, left, or right), count (two to five balls), digit (one to nine), and position (a ball on the left or the right). In every family each value is pushed toward the next value in a fixed cyclic order, giving eight runs per cell (9 for digits).

Capability gate. A family is analysed on a model only if the judge names the prompted value on at least 0.75 of that model's unintervened images of the family. Families that fail are reported with their accuracy and not analysed, because an intervention cannot be scored on a value the model does not draw recognisably when asked to.

Depths. Layers are specified by relative depth and mapped to each model: a nine-point grid for the sweep (1B: L0, L3, L6, L9, L12, L15, L18, L21, L23; 7B: L0, L4, L8, L11, L15, L19, L23, L26, L29), five depths at $0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ for identity and properties, the three middle depths of those for swap and ablation, three early depths for clamping, and the band L6–L18 (1B) or L8–L23 (7B) for multi-layer steering.

Directions. Every arm writes one unit-norm direction per prompt position, scaled to the norm of the activation it is added to. $\texttt{dm}$ is the activation of the target prompt minus the mean over the other prompts of the same set (the 24 concepts, or the values of the family), computed per position. $\texttt{rot}$, the rotated control, is the same direction under a fixed random orthogonal matrix Q. $\texttt{lens}_{\text{text}}$ and $\texttt{lens}_{\text{gen}}$ are the target token's J-lens row, $(W_U \operatorname{diag}(g)\, J_\ell)_w$, from the fit on pretraining text and from the fit in the image-generation context, both at $n = 1000$. $\texttt{lens}_{\text{rot}}$ is the text row under the same Q. This separates the row's content from the form "one vector repeated at every position". The decomposition arms split $\texttt{dm}$ by sparse non-negative pursuit over the whole text dictionary (102,400 atoms, $K = 16$). $\texttt{dm}_J$ is the reconstruction, its component inside the text part of the MJ-space, and $\texttt{dm}_{\neg J}$ is the remainder, so that the two sum to $\texttt{dm}$. All arms act on the same prompt positions, from the first token at which the reference prompts differ to the end of the prompt. The multi-layer experiment also steers during the rollout.

Operations. The four interventions of the reference work are run: additive steering, $h \leftarrow h + \alpha \|h\| u$; swap in the basis of the two directions, with coordinates $c = (V^\top V)^{-1} V^\top h$ and $h \leftarrow h + \alpha V(\tilde{c} - c)$ where $\tilde{c}$ exchanges the two coordinates; ablation, projecting the host direction out; and clamping, adding $\texttt{dm}$ while the target's text-lens coordinate is held at its value in the unintervened pass at every later layer. By working point we mean the step size and the layer at which the intervention is applied.

Working point. The step size α^* is fixed from the rotated control of the sweep alone (Table 26), as the largest α at which that arm keeps image integrity in at least two thirds of runs at every layer. At 7B this gives $\alpha^* = 0.5$. At 1B no α meets the rule at every layer, so there $\alpha^* = 0.5$ is the largest α that meets it pooled over layers. Identity and clamping are also run at $2\alpha^*$; these runs were added after the sweep was seen and are reported as a sensitivity analysis.

Judge and criteria. All images are scored by Qwen2.5-VL-7B-Instruct, which generates none of them, on first-token logit margins. Identity is installed in a run when the judge prefers the target to the host ($m_{\text{id}} > 0$) and the image is intact. A property is written when the judge's choice among the family's values is the target and the image is intact. A host is removed by ablation when the judge answers that it is absent and the image is intact. Integrity is the judge's preference for "photograph" over "noise" ($m_{\text{int}} > 0$), or for "picture" over "noise" for five of the six drawn families (§I.2). Each question is scored in its own forward pass. Rates carry Wilson 95% intervals, and arms are compared on matched runs by an exact McNemar test. The protocol's test of its central claim, that lens directions cannot control generation, is stated in advance: lens directions control generation on a model if, in any cell, a lens arm succeeds in at least half the runs with a Wilson lower bound above the rotated control's rate in the same cell, and if that happens the claim is withdrawn for that model and family.

I.2 Deviations from the written protocol

Two points departed from the written protocol. First, the drawn families (size, shape, direction, count, digit, position) may be scored for integrity by the judge's preference for "picture" over "noise" rather than "photograph" over "noise". Both questions are asked of every non-identity image, and a drawn family uses the "picture" question when the judge's median "photograph" margin on that family's unintervened images is not positive. This selects "picture" for shape, direction, count, digit, and position; size keeps "photograph". Second, the 7B decomposition arms use the $n = 400$ checkpoint of the full $J^{T \to T}$ fit rather than $n = 1000$; the fit was continued to $n = 1000$ for the calibration report only, and the 1B decomposition uses $n = 1000$. Both points affect reproduction and are reported here.

I.3 CALIBRATION OF THE LENS

Every lens row the protocol writes is estimated by one vector–Jacobian product per prompt, with the row's cotangent placed on every target position. This is the estimator of the full operator restricted to the rows needed. On 150 Janus-Pro-1B prompts it reproduces the rows of the full operator with a minimum cosine of 0.99979. Rows are accumulated in cumulative checkpoints and in odd and even halves. The adequacy rule of the protocol is that a lens-arm null at a layer is interpreted only if the median split-half cosine of the rows at $n = 1000$ is at least 0.95 there (Table 22).

Table 22: Calibration of the lens rows used in the writing experiments (51 tokens, or 32 codebook entries for the rollout operators $J^{T\to V}$ and $J^{V\to V}$; every layer). For each model, context and operator: the split-half cosine of the weakest layer at four sample sizes, the median layer at $n = 1000$, and the layers that stay below the adequacy threshold of 0.95 at $n = 1000$.

model	context	operator	weakest layer, split-half cosine at $n =$				median	layers < 0.95
			10	50	200	1000	at 1000	at n=1000
Janus-Pro-1B	text	$J^{T\to T}$	0.499	0.915	0.977	0.995	0.9998	none
	generation prompt	$J^{T\to T}$	0.336	0.802	0.985	0.985	0.9999	none
	generation rollout	$J^{T\to V}$	0.159	0.446	0.912	0.983	0.9994	none
	generation rollout	$J^{V\to V}$	0.800	0.966	0.993	0.998	0.9994	none
	understanding	$J^{T\to T}$	0.966	0.992	0.999	1.000	1.0000	none
	understanding	$J^{V\to T}$	0.906	0.978	0.996	0.999	0.9997	none
Janus-Pro-7B	text	$J^{T\to T}$	0.388	0.752	0.927	0.984	0.9995	none
	generation prompt	$J^{T\to T}$	0.622	0.954	0.991	0.997	0.9997	none
Qwen2.5-VL-7B	text	$J^{T\to T}$	0.638	0.834	0.963	0.989	0.9984	none
	understanding	$J^{T\to T}$	0.924	0.983	0.997	1.000	0.9996	none
	understanding	$J^{V\to T}$	0.678	0.920	0.982	0.996	0.9983	none
Qwen3-VL-32B	text	$J^{T\to T}$	0.041	0.265	0.507	0.721	0.9922	L0–L3
	understanding	$J^{T\to T}$	0.933	0.956	0.919	0.995	0.9990	none
	understanding	$J^{V\to T}$	0.148	0.375	0.435	0.770	0.9812	L0–L26

Every row of both Janus models and of Qwen2.5-VL-7B clears the threshold at every layer by $n = 1000$, the lowest value being 0.983, at L1 of the Janus-Pro-1B rollout rows $J^{T\to V}$. Every lens-arm null in §3.5 is therefore interpretable under the protocol. The weakest layer is what needs the sample size: for the Janus-Pro-7B text rows it is at 0.752 with 50 prompts and 0.862 with 100, while the median layer is above 0.99 from $n = 50$ on. Qwen3-VL-32B is the exception: its $J^{V\to T}$ rows stay below the threshold at L0–L26 of 64 layers even at $n = 1000$, rising from 0.770 at L0, and its text-context rows stay below it at L0–L3. Qwen3-VL-32B's $J^{V\to T}$ is under-calibrated through the first two fifths of the network, so its operator comparison is reported as undetermined (Appendix F.3).

Table 23: Split-half agreement of the full operators: linear CKA between the estimates from odd and even prompts, at the weakest and the median layer.

fit	operator	prompts	CKA, weakest layer	CKA, median layer
Janus-Pro-1B, pretraining text	$J^{T\to T}$	1000	0.9993	0.9999
Janus-Pro-7B, pretraining text	$J^{T\to T}$	400	0.9936	0.9996
Janus-Pro-7B, pretraining text	$J^{T\to T}$	1000	0.9970	0.9997
Janus-Pro-1B, text-output context	$J^{T\to T}$	150	0.989	1.000
Janus-Pro-1B, text-output context	$J^{V\to T}$	150	0.997	0.998
Janus-Pro-1B, image-output context	$J^{T\to V}$	150	0.965	1.000
Janus-Pro-1B, image-output context	$J^{V\to V}$	150	0.992	1.000
Qwen2.5-VL-7B-Instruct	$J^{T\to T}$	240	0.996	0.999
Qwen2.5-VL-7B-Instruct	$J^{V\to T}$	240	0.981	0.991
Qwen3-VL-32B	$J^{T\to T}$	60	0.250	0.928
Qwen3-VL-32B	$J^{V\to T}$	60	0.006	0.076

The full text operators of the decomposition arms are adequately estimated at both sizes (CKA at least 0.9936 at the weakest layer of the $n = 400$ 7B fit). The 60-prompt full operators of Qwen3-VL-32B are not: their $J^{V\to T}$ has a median split-half CKA of 0.076 (Table 23).

I.4 FULL RESULTS

The failures of the lens arms are of a different kind from those of the controls (Table 24). At the same step, $\texttt{lens}_{\text{text}}$ and $\texttt{lens}_{\text{gen}}$ destroy the image in 456 and 587 of 720 runs at 7B, against 241

Table 24: Outcome of every identity run by arm: installed; image intact but the target not installed; image integrity lost.

	dm	$\texttt{dm}_{\neg J}$	$\texttt{dm}_{J}$	$\texttt{lens}_{\text{text}}$	$\texttt{lens}_{\text{gen}}$	$\texttt{lens}_{\text{rot}}$	rot
Janus-Pro-7B, identity, 720 runs per arm							
installed	127	124	3	0	1	2	3
intact, not installed	439	418	452	264	132	471	476
integrity lost	154	178	265	456	587	247	241
Janus-Pro-1B, identity, 720 runs per arm							
installed	67	61	10	8	4	7	16
intact, not installed	406	403	273	141	175	376	393
integrity lost	247	256	437	571	541	337	311

for the rotated control and 247 for $\texttt{lens}_{\text{rot}}$, while installing the target in 0 and 1; dm, at the same step, installs it in 127. The same ordering holds at 1B. The step is therefore strong enough to act on the image, and the lens rows fail because they are not usable as write directions, not because the intervention is too weak.

Table 25: Identity: runs installing the target, out of 72 per cell (24 pairs $\times$ 3 templates), at $\alpha^* = 0.5$ and $2\alpha^* = 1.0$.

model	α	layer	dm	$\texttt{dm}_{\neg J}$	$\texttt{dm}_{J}$	$\texttt{lens}_{\text{text}}$	$\texttt{lens}_{\text{gen}}$	$\texttt{lens}_{\text{rot}}$	rot
Janus-Pro-7B	0.5	L0	0	0	0	0	0	0	0
		L8	29	25	2	0	0	0	0
		L15	2	1	0	0	1	1	0
		L23	1	0	0	0	0	0	1
		L29	0	0	0	0	0	0	0
	1	L0	41	40	0	0	0	0	1
		L8	43	46	1	0	0	0	1
		L15	10	8	0	0	0	0	0
		L23	1	2	0	0	0	1	0
		L29	0	2	0	0	0	0	0
Janus-Pro-1B	0.5	L0	3	4	2	1	0	0	2
		L6	4	1	3	4	1	0	0
		L12	1	1	0	1	0	3	1
		L18	1	1	0	0	1	1	3
		L23	1	1	3	1	1	1	3
	1	L0	35	31	0	0	0	0	5
		L6	16	16	1	1	0	0	0
		L12	3	3	0	0	0	2	2
		L18	2	0	0	0	0	0	0
		L23	1	3	1	0	1	0	0

The decomposition uses the text lens dictionary, so the MJ component is the part of the MJ-space spanned by the text dictionary, not the MJ-space as a whole. At 7B that component accounts for 47% of the direction's variance at every middle layer and 95% at L0, yet installs the target in 3 of 720 runs, while the remainder installs it in 124, as often as the whole direction (Tables 25 and 29). At L0 the remainder is 5% of the direction and, written at the same norm as the whole, installs the target in 40 of 72 runs at $2\alpha^*$ (Figure 5a). Janus-Pro-1B shows the same pattern at lower rates (Figure 5b).

The reference work's own operations do not write either when they are performed with the lens rows (Table 27). At 7B, swapping the host's and the target's coordinates installs the target in 24 of 24 runs at L8 in the basis of their difference-of-means directions, and in at most 2 of 24 runs in any cell in the basis of their lens rows. Projecting out the host's difference-of-means direction removes the host in 5 of 24 runs at L8; projecting out its two lens rows or the two rotated controls removes it in none of 288 runs. Clamping the target's lens coordinate while dm is added does not prevent the installation (76 against 83 of 144 runs, $p = 0.25$), so the direction's effect does not pass through the coordinate the lens assigns to the target.

Multi-layer steering is uninformative at this working point. Every arm, at both sizes, both step sizes, and both position sets, installs the target in 0 of 24 runs, and every one of those 960 images fails the integrity check. Adding the direction at all 13 (1B) or 16 (7B) layers of the band at once sums the displacement over the band, and a step the rotated control survives at one layer destroys the image when applied at sixteen. The experiment therefore says nothing about any arm and is reported only for completeness.

On the four property families that pass the gate at 7B, dm writes the target value in 129 of 320 runs, against 12 for $\texttt{lens}_{\text{text}}$, 4 for $\texttt{lens}_{\text{gen}}$, and 14 for the rotated control (Table 28, Figure 5c).

Table 26: The sweep: two pairs (*dog→apple*, *car→apple*) × two seeds, template T1; runs installing the target, out of 4 per cell. Janus-Pro-7B above, Janus-Pro-1B below.

arm	α	L0	L4	L8	L11	L15	L19	L23	L26	L29
dm	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	4	4	0	0	0	0	0	0
	2	4	0	0	3	1	0	0	0	0
rot	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	1	0	0	0	0	0	0
$\text{lens}_{\text{text}}$	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
lens_{gen}	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
lens_{rot}	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0

arm	α	L0	L3	L6	L9	L12	L15	L18	L21	L23
dm	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0	0
rot	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
$\text{lens}_{\text{text}}$	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
lens_{gen}	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
lens_{rot}	0.25	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0

Taken together, on Janus-Pro the Jacobian-lens rows are not suitable as write directions: difference-of-means directions can install a concept, lens rows cannot. This shows that a readout lens does not automatically provide control directions; the limitation these results establish is that of the lens rows as write directions, and not necessarily of the MJ-space. One cell meets the withdrawal criterion (Janus-Pro-1B, position family, L0, $\alpha = 1$: $\text{lens}_{\text{text}}$ succeeds in 4 of 8 runs against 0 for rot), so by the rule fixed in advance the claim is withdrawn for that model and family. Its images show that the criterion should not be read as a successful installation, since only one of its eight images contains a ball, so at least three of the four runs the judge scores as successes contain none.

J LAYER PROFILES, REPLICATION, AND THE EFFECT OF MODEL SCALE

This appendix supplements §3.5. It reports measurements made before the protocol of Appendix I. These measurements cover two things: the depth at which each modality becomes decodable, and the identity and colour profiles of a separate, smaller protocol at seven layers of Janus-Pro-1B and six of Janus-Pro-7B, together with their replication. The separate protocol scored identity with the model's own two-alternative judge and colour with pixel opponent channels. Its profiles agree with those of the later protocol where the two overlap. We report it because the colour-specificity measurement has no counterpart in the later protocol.

Table 27: The reference work's other operations: coordinate swap and ablation at the three middle depths, and clamping the target's lens coordinate while adding dm at three early depths. Successes out of 24 runs per cell (one per pair, template T1).

			Janus-Pro-7B			Janus-Pro-1B		
experiment	arm	α	L8	L15	L23	L6	L12	L18
swap	dm	1	24	8	0	10	0	2
	dm	2	20	12	0	6	0	1
	$\texttt{lens}_{\text{text}}$	1	1	0	0	2	1	0
	$\texttt{lens}_{\text{text}}$	2	2	1	0	3	1	1
	$\texttt{lens}_{\text{gen}}$	1	0	0	0	2	1	0
	$\texttt{lens}_{\text{gen}}$	2	1	0	0	2	2	3
ablate	dm	—	5	1	0	6	2	2
	rot	—	0	0	0	2	2	3
	$\texttt{lens}_{\text{text}}$	—	0	0	0	2	2	2
	$\texttt{lens}_{\text{gen}}$	—	0	0	0	4	3	3
	$\texttt{lens}_{\text{rot}}$	—	0	0	0	2	2	4
			L4	L8	L11	L3	L6	L9
clamp	dm	0.5	14	12	2	3	3	3
	dm	1	19	23	13	10	6	5
	dm+clamp	0.5	14	13	1	3	5	2
	dm+clamp	1	19	22	7	7	5	3

Table 28: Property families, pooled over five depths and two step sizes (80 runs per entry, 90 for digits). The gate column gives the judge's accuracy on the unintervened images; families marked $^{\times}$ fail the gate and are not analysed.

model	family	gate	dm	rot	$\texttt{lens}_{\text{text}}$	$\texttt{lens}_{\text{gen}}$	$\texttt{lens}_{\text{rot}}$
Janus-Pro-7B	colour	8/8	39/80	0/80	0/80	1/80	0/80
	direction	6/8	31/80	12/80	8/80	1/80	13/80
	position	8/8	33/80	2/80	2/80	2/80	0/80
	shape	8/8	26/80	0/80	2/80	0/80	2/80
	count	5/8$^{\times}$	26/80	11/80	5/80	1/80	19/80
	digit	3/9$^{\times}$	20/90	4/90	1/90	2/90	4/90
	size	4/8$^{\times}$	14/80	18/80	6/80	0/80	15/80
Janus-Pro-1B	colour	7/8	23/80	3/80	3/80	1/80	3/80
	position	8/8	43/80	4/80	8/80	6/80	6/80
	shape	8/8	15/80	0/80	2/80	0/80	0/80
	count	2/8$^{\times}$	22/80	25/80	6/80	17/80	15/80
	digit	2/9$^{\times}$	5/90	4/90	0/90	1/90	0/90
	direction	4/8$^{\times}$	15/80	13/80	6/80	8/80	15/80
	size	4/8$^{\times}$	14/80	18/80	5/80	6/80	15/80

J.1 Where each modality becomes decodable

The landmark is the first layer at which a logit lens applied to the residual stream recovers the model's own *next* emitted token on a given fraction of positions. Scoring against the next token rather than the current token matters. At a visual position, the residual stream begins as the embedding of the token sitting there. Scoring against the current token therefore measures the input echo of §3.4 and saturates immediately. We report quartile landmarks rather than accuracies because the two modalities reach different ceilings, and a landmark normalises that away.

model	positions	n	first layer reaching				depth
			25%	50%	75%	90%	
Janus-Pro-1B	text	640	17	20	23	24	24
Janus-Pro-1B	image	11500	21	23	24	24	24
Janus-Pro-7B	text	512	19	25	29	30	30
Janus-Pro-7B	image	9200	28	30	30	30	30

Both models place the image later than the text, and the gap is wider in the deeper model. Four layers separate the 25% landmarks at 1B, against nine at 7B. The two models have different depths, 24 and 30 layers, so the relevant quantity is the fraction of depth rather than the layer index. On that scale, the two landmarks move in opposite directions as the model grows. Text becomes decodable earlier: 0.71 of depth at 1B against 0.63 at 7B. The image becomes decodable later: 0.88 against 0.93. The gap widens because both landmarks move, not because one modality shifts on its own, so the deeper model is not merely a rescaling of the shallower one.

Table 29: Paired comparisons on matched runs: runs in which only the first arm succeeds, runs in which only the second does, and the exact McNemar p. The property rows are restricted to the families that pass the gate.

		Janus-Pro-7B			Janus-Pro-1B		
experiment	comparison (first vs second)	only 1st	only 2nd	p	only 1st	only 2nd	p
identity	dm vs rot	125	1	3×10^{-36}	61	10	5×10^{-10}
identity	dm vs $\text{lens}_{\text{text}}$	127	0	1×10^{-38}	62	3	2×10^{-15}
identity	dm vs lens_{gen}	126	0	2×10^{-38}	66	3	2×10^{-16}
identity	$\text{dm}_{\neg J}$ vs dm_{J}	121	0	8×10^{-37}	58	7	4×10^{-11}
identity	$\text{dm}_{\neg J}$ vs dm	22	25	0.77	23	29	0.49
identity	$\text{lens}_{\text{text}}$ vs rot	0	3	0.25	8	16	0.15
identity	$\text{lens}_{\text{text}}$ vs lens_{rot}	0	2	0.50	8	7	1.00
swap	dm vs $\text{lens}_{\text{text}}$	61	1	3×10^{-17}	15	4	0.02
swap	dm vs lens_{gen}	63	0	2×10^{-19}	17	8	0.11
ablate	dm vs rot	6	0	0.03	9	6	0.61
clamp	dm vs dm+clamp	17	10	0.25	16	11	0.44
properties	dm vs rot	122	7	3×10^{-28}	77	3	1×10^{-19}
properties	dm vs $\text{lens}_{\text{text}}$	122	5	3×10^{-30}	72	4	4×10^{-17}
properties	$\text{lens}_{\text{text}}$ vs rot	9	11	0.82	11	5	0.21
properties	lens_{gen} vs rot	4	14	0.03	4	4	1.00

J.2 IDENTITY BY LAYER

The intervention pushes a difference-of-means direction for one concept into a prompt naming another. It acts at all prompt positions during prefill, regenerates from the same seed, and scores the decoded image with the model's own two-alternative judge. A margin of zero is chance. By relative depth we mean the layer index divided by the total number of layers.

layer	α=0	dm	lens	rot	dm−rot
Janus-Pro-1B (24 blocks)					
L0	-6.88	+3.38	+1.25	-5.50	**+8.88**
L4	-7.12	+2.00	+0.25	+0.75	**+1.25**
L8	-7.00	+0.75	+0.00	-5.75	**+6.50**
L12	-7.12	-1.38	+0.50	-4.12	**+2.75**
L16	-7.00	-6.50	+0.50	-5.88	**-0.62**
L20	-7.00	-6.00	+1.50	-4.62	**-1.38**
L23	-7.00	-7.75	—	-7.38	**-0.38**
Janus-Pro-7B (30 blocks)					
L0	-8.88	+12.88	—	-8.97	**+21.84**
L8	-8.94	+13.50	—	+1.50	**+12.00**
L16	-8.94	+0.38	—	-9.66	**+10.03**
L20	-8.62	-9.34	—	-8.44	**-0.91**
L24	-9.22	-10.97	—	-10.00	**-0.97**
L28	-8.75	-9.09	—	-9.06	**-0.03**

The two models behave differently, and not by degree. At 7B the direction installs the pushed object at L8, a quarter of the way into the network, with a margin that is both large and confidently signed. At 1B nothing comparable happens at any depth. The best cell is L0, which is close to editing the prompt embedding and therefore carries little interpretive weight. By L12 the margin is negative again, and by L16 it has returned to the unintervened value. Reading the 1B profile alone would support the conclusion that object identity cannot be installed in the image by a direction added to the residual stream. That conclusion is wrong, and it is wrong in the direction that would have been convenient. This is why the main text reports the identity result at the larger model rather than the smaller one.

The lens column of this protocol was run at 1B only. It added the lens row at every prompt position, including the start token and the template tokens before the concept word. The difference-of-means directions leave those positions exactly untouched, because they are identical in every prompt of the set. The lens entries therefore differ from the other columns in where the intervention acts and are not comparable with them. Their margins near zero come from damaged images, and with the positions differing this protocol cannot attribute that damage to the direction itself. The lens arms of

the later protocol act on the same positions as every other arm and still damage the image far more often than a rotated direction (Appendix I), so the damage is not explained by position.

J.3 Appearance by layer

At each layer we push each of four colour terms in turn and measure the four corresponding pixel opponent channels, as changes from the unintervened image. The diagonal entry is the channel matching the term pushed. The table gives the four diagonal entries, the means of the diagonal and the off-diagonal entries, and the mean of the rotated control. The four channels are not independent. By construction the yellow channel is the negative of the blue one, and the red, green and blue channels sum to zero, so the four span two independent axes. The diagonal mean is therefore the summary to read, and the individual entries are shown for completeness. By appearance we mean colour, direction, position, and shape; this protocol measures colour only.

layer	diagonal entry				mean	mean	mean	diag
	red	green	blue	yellow	diag	off	rot	− rot
Janus-Pro-1B								
L0	+2.6	+4.4	-0.1	+16.0	+5.7	-1.7	-0.6	**+6.3**
L4	+7.0	+1.3	+2.8	-14.9	-0.9	-0.8	+0.9	**-1.9**
L8	-7.3	+3.2	+23.1	+17.0	+9.0	-4.7	-0.2	**+9.2**
L12	-5.1	+7.4	+22.9	+4.3	+7.4	-6.4	-4.1	**+11.5**
L16	+15.6	+13.4	+27.3	+17.0	+18.3	-7.1	-0.8	**+19.2**
L20	+18.9	+15.0	+15.6	-3.3	+11.6	-4.4	-5.7	**+17.2**
L23	-1.6	+1.7	+7.7	-5.3	+0.6	-2.1	+2.0	**-1.4**
Janus-Pro-7B								
L0	-0.1	-0.1	+1.3	-0.9	+0.1	-0.8	-1.0	**+1.1**
L8	+23.5	+1.9	+33.9	+9.2	+17.1	-4.7	-0.7	**+17.8**
L16	+17.3	+11.5	+24.7	+16.9	+17.6	-6.3	+0.1	**+17.5**
L20	+17.9	+20.1	+26.6	+10.2	+18.7	-6.4	+0.5	**+18.2**
L24	+46.4	+7.6	+23.0	-8.6	+17.1	-5.2	+1.1	**+16.1**
L28	+15.2	-9.5	+10.8	-2.6	+3.5	-2.4	+0.3	**+3.2**

In both models the difference-of-means direction installs the pushed colour across a broad middle band, and at the deepest layer measured, L23 at 1B and L28 at 7B, the effect is small or absent. The appearance result, unlike the identity one, therefore does not depend on which model is examined. The band is wider at 7B in absolute layers and comparable as a fraction of depth.

J.4 Replication

The colour runs below vary the random seed and the subject. Across the nine runs inside the middle band, 35/36 diagonal entries are positive; the run at L28 of 7B repeats the deepest layer, outside the band, and is left out of that count. The four within-family directions sum to zero by construction, so entries within a run are not independent. We take the run as the unit of replication.

model	subject	seed	diagonal entry				mean	mean
			red	green	blue	yellow	diag	rot
1B L8–16	dog	0	+47.9	+60.8	+45.8	+31.2	+46.4	-3.4
1B L8–16	dog	1	+18.0	+43.5	+41.1	+3.9	+26.6	-2.0
1B L8–16	dog	2	+14.2	+84.8	+68.2	+42.9	+52.5	-3.8
1B L8–16	cat	3	+31.6	+15.5	+81.9	+46.1	+43.8	+4.0
1B L8–16	cat	4	+21.7	+22.8	+66.7	+70.4	+45.4	+1.1
7B L16	dog	1	+5.1	+5.9	+29.1	+18.8	+14.7	+0.5
7B L16	dog	2	+8.8	+8.2	+33.1	+9.7	+15.0	-0.8
7B L16	cat	3	-1.4	+7.5	+29.4	+15.7	+12.8	-1.3
7B L28	dog	1	+3.8	-7.4	+4.5	+6.7	+1.9	-1.6
7B L8	dog	1	+9.9	+10.8	+34.1	+18.6	+18.4	+2.4

The identity result at 7B is replicated by varying the seed, the subject, and the pushed concept, holding the layer at L8 where the main-text effect is largest; the runs at deeper layers vary the seed

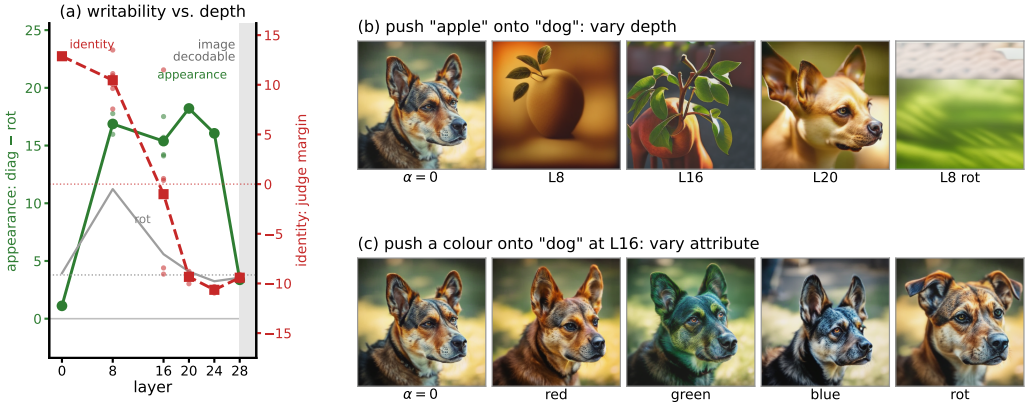

Figure 13: **A difference-of-means direction writes appearance across the middle band and identity only at early layers.** Janus-Pro-7B, difference-of-means directions, images decoded by the model's own VQ decoder, scored by the model's own judge and by pixel opponent channels (Appendix J). *(a)* Appearance against the rotated control (left axis, opponent-channel units) and the identity judge's margin (right axis, dotted lines at chance and at the unintervened value), with individual runs as dots; the units differ, so what the panel compares is the shapes. *(b)* Identity at fixed concept and varying depth. L8 installs the pushed object in all six runs; L16 varies by seed between installing and doing nothing, the cell shown being one that carries the apple's colour, stem and leaves on the dog's body; L20 leaves an undamaged dog. The rotated control destroys the image, and the judge scores texture at chance—the reason the pictures are here and not only the numbers. *(c)* Appearance at fixed depth and varying attribute: the subject survives, the colour is the one pushed, and the rotated control changes the sample without changing its colour.

alone. At L8 the judge favours the pushed object in all five replications, as it does in the original run. This shows that the 1B null does not generalize to the larger model.

layer	subject	pushed	seed	α=0	dm	rot	dm−rot
L16	dog	apple	1	-9.12	+11.50	+1.25	**+10.25**
L16	dog	apple	2	-9.44	-8.44	-9.12	**+0.69**
L16	dog	apple	3	-9.03	+0.56	-8.62	**+9.19**
L16	dog	apple	4	-9.59	-9.06	-9.12	**+0.06**
L20	dog	apple	1	-9.09	-10.03	-8.31	**-1.72**
L20	dog	apple	2	-9.28	-8.78	-8.97	**+0.19**
L20	dog	apple	3	-9.12	-9.19	-9.50	**+0.31**
L24	dog	apple	1	-8.97	-10.31	-9.56	**-0.75**
L28	dog	apple	1	-8.94	-9.75	-9.78	**+0.03**
L8	dog	apple	1	-9.06	+10.91	+1.50	**+9.41**
L8	dog	apple	2	-9.16	+11.12	-0.38	**+11.50**
L8	car	apple	3	-7.31	+10.03	+1.88	**+8.16**
L8	dog	flower	4	-15.00	+9.62	-3.38	**+13.00**
L8	clock	apple	5	-6.44	+7.56	-4.12	**+11.69**

J.5 Colour specificity

At 7B, across L8–L24, the pushed colour exceeds the rotated control by 15–18 opponent units on the diagonal, averaged over the runs at each layer, and the subject survives (Figure 13). Over the nine replication runs inside the middle band, five at 1B and four at 7B, 35 of 36 diagonal entries are positive. The later protocol, whose judge names the colour it sees, gives the same result without resolving this specificity.

J.6 What the scale comparison does and does not license

Two models are enough to show that the 1B null does not generalize to unified models, and not enough to establish a trend in model size.