THE READOUT DOES NOT ROTATE: REFERENCE-FRAME BINDING FAILURE IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models frequently err in spatial perception when direction is defined by a reference object's heading. To investigate the cause, we fix the prompt, the target, and the camera position in controlled synthetic scenes, reverse only the reference object's heading, and examine eight open-source models stage by stage across perception, transmission, and answering. We find that both the reference-object heading and the camera information are linearly decodable and are carried to the answer token, yet the heading information is not used, and the model preferentially relies on the camera information instead. We then attempt a targeted fine-tuning remedy, but find that the model only memorizes the third-party headings it has seen and does not generalize what it has learned. Taken together, we conclude that the failure of perception in a third-party coordinate frame occurs in the mapping from representations to answers, and that fine-tuning does not repair this defect. We propose a set of metrics as a quantitative standard for this ability, which can test whether a proposed remedy genuinely establishes reference-frame binding and thereby support the evaluation of future work.

1 INTRODUCTION

Vision-language models (VLMs) still make errors on basic spatial questions involving just two objects (Kamath et al., 2023; Wang et al., 2024; Tong et al., 2024). This paper focuses on *reference-frame binding*, that is, specifying a reference frame for the model to use when judging spatial relations. For example, "The ball is left of the car" can be interpreted relative to either the observer or the car, and changing the reference frame may change the answer. The former is a relative, observer-centered *egocentric frame*, while the latter is an intrinsic, object-centered *allocentric frame* (Levinson, 2003; 1996). To answer in the latter frame, the model must not only recognize the car and locate the ball but also bind the reference frame to the car.

In embodied manipulation and navigation, this distinction determines how a spatial instruction is translated into a target location. A command to approach a car from its own left side can specify the camera's right. Binding errors can therefore misdirect a downstream decision despite correct object localization, motivating tests of whether the requested frame actually governs the answer.

Consider asking the same question about a pair of images that differ only in the reference object's heading and have opposite correct answers. If the model gives the same answer to both images in a pair, its overall accuracy should be 0.5, but this does not reveal where the model fails. Because overall accuracy conflates these behaviors, we measure changes in the model's predictions under heading reversal. The relevant distinction is whether the answer changes with the reference object or stays tied to the camera.

Behavioral benchmarks document reference-frame difficulties and preferences for egocentric reference frames (Zhang et al., 2025b; 2026). A model may fail to represent heading, lose it before answering, or retain it without using it appropriately. These failures call for improvements in perception, transmission, and answer mapping, respectively. We therefore test the three stages in one controlled setting, comparing heading with camera information.

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.

Experimental results show that heading-related visual information is decodable at image and answer positions. Reflecting its naturally encoded component changes answers only weakly, whereas replacing camera-related image activations has a strong effect. The answer function hardly rotates when the reference object reverses. After fine-tuning, models fit the training data and generalize across objects but produce almost exactly opposite answers at an unseen heading. The failure lies in turning available information into an answer in the requested frame.

We make three contributions.

- We combine behavioral tests, representational probes, and causal interventions to examine the perception, transmission, and answer mapping of reference-object heading information stage by stage. We find that heading information is decodable and reaches the answer position, yet has little influence on spatial judgments, while camera information has a stronger effect on answers. These results support localizing reference-frame binding failure to the mapping from representations to answers.
- We propose a comprehensive set of metrics for reference-frame binding, combining paired correctness, phase rotation, and antipodal errors to provide a quantitative standard for evaluating whether existing model improvement strategies establish this ability.
- Through targeted fine-tuning under different configurations, we separate object generalization from heading generalization, revealing a directional rule that transfers across objects but fails under heading reversal.

2 Related Work

Linguistic and cognitive accounts distinguish relative, intrinsic, and absolute frames of reference (Levinson, 2003; Frank, 1998; Logan & Sadler, 1996). COMFORT and SpinBench bring frame ambiguity, consistency, and viewpoint transformations into VLM evaluation (Zhang et al., 2025b; 2026). We build on their behavioral observations by using heading-reversal pairs to test the internal stages that could produce the errors.

SpatialVLM uses spatial predictions to annotate rewards for robot manipulation (Chen et al., 2024), while VSI-Bench tests spatial memory and reasoning from videos of indoor environments (Yang et al., 2025). These applications and evaluations motivate checking whether the coordinates specified by an instruction govern spatial judgments. Our paired tests can complement localization and scene-understanding evaluations by measuring whether answers change with the reference object's heading.

Recent mechanistic studies distinguish spatial encoding from its use. Kang et al. (2026) find linear spatial IDs that bind image-plane positions to object-word activations and causally affect answers. In a text-only perspective-rotation task, Yang et al. (2026) observe that viewpoint information can be encoded without being bound to the corresponding observation. Object-centered judgments require image-plane positions to be interpreted using the reference object's heading. We examine whether VLMs use a visually supplied heading cue to perform this transformation.

Depth intermediates, scene graphs, captions, and explicit reasoning have been proposed to improve spatial reasoning (Chen et al., 2024; Liu et al., 2025; 2026; Weng et al., 2025). We use probes and activation replacement to measure which information reaches the answer position and which information changes the answer (Hewitt & Liang, 2019; Meng et al., 2022; Zhang & Nanda, 2024). Heading-reversal tests then measure whether answers follow the requested coordinate frame. Appendix N gives the broader comparison.

3 Experimental Design

Most experiments use Blender-rendered COMFORT scenes (Zhang et al., 2025b). A target object orbits a reference object in 10° steps. Each of two opposite headings has 37 sampled entries, including a repeated endpoint, across 10 reference-object configurations. We denote the target's camera-relative azimuth by θ. In image pairs where the reference object's heading is reversed, the target object's position and the camera viewpoint remain unchanged. Camera-reversal pairs rotate the camera by 180° while preserving the objects and reference heading. These settings isolate two different

changes in visual evidence. Appendix A provides detailed behavioral tests of the models' failures on allocentric questions, and Appendix E describes the specific settings of the experiments outlined above.

Questions provided to the models explicitly distinguish egocentric and allocentric reference frames, for example, "From your viewpoint, is the ball left of the car?" versus "Taking the car's front as forward, is the ball to its left?" Most experiments use yes/no questions. The crossed fine-tuning experiment uses four choices: left, right, front, and back. We test eight open VLMs spanning six language-model families, each with at most 8B parameters. Their full names and backbones appear in Appendix C.1. The primary models for detailed causal interventions, image-input controls, and initial fine-tuning are InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, all with Qwen-family language backbones.

For the subsequent probe experiments, we assign all samples associated with the same image together to either the training set or the test set. We also conduct a series of control experiments to prevent spurious results caused by data leakage from misleading our interpretation. For causal tests, we fit subspaces on training examples and evaluate interventions on separate test examples. Probes measure recoverable information; interventions measure its effect on the model's answer.

4 Behavior Under Heading Reversal

For each image pair, we ask the same allocentric question. Let L denote the image in the pair where the reference object faces left, and R the image where it faces right. Let $p_L^{(i)}$ and $p_R^{(i)}$ be the probabilities of answering yes, and let $t_L^{(i)}, t_R^{(i)} \in \{0, 1\}$ be complementary correct labels. For P image pairs, with $\eta_i = \mathrm{sign}(t_L^{(i)} - t_R^{(i)})$, define directed correctness D and sign agreement (SA) as

$$D = \frac{1}{P}\sum_{i=1}^{P} \eta_i \big(p_L^{(i)} - p_R^{(i)}\big), \qquad \mathrm{SA} = \frac{1}{P}\sum_{i} \mathbf{1}\left[\mathrm{sign}(p_L^{(i)} - p_R^{(i)}) = \eta_i\right]. \tag{1}$$

D records both the direction and magnitude of the response to heading. It is zero for invariant probabilities and lies in $[-1, 1]$; ideal predictions $p_L = t_L, p_R = t_R$ give $D = 1$. SA records only whether the probability change has the correct sign. Full results are in Appendix A.2.

Heading reversal elicits almost no correctly directed response on COMFORT. Across the four tested models, $|D|$ never exceeds 0.004, SA stays at or below 0.500, and accuracy on allocentric questions remains near chance (Table 2). In tests on real-world scenes in SpinBench (Zhang et al., 2026), we observe a preference for the egocentric perspective among incorrect model responses. For most incorrectly answered allocentric questions with an identifiable egocentric answer, the same response would be correct if the question were changed to an egocentric question (Table 4).

5 Tracing the Mechanism

The behavioral preference could originate in visual encoding, transmission to the answer position, or the final use of that information. Probes test the first two stages; interventions compare the answer effects of heading and camera signals at matched sites. Figure 1 shows the resulting mechanism in InternVL3-8B.

5.1 Decoding heading from image representations

For the perception stage, we decode heading information from the residual stream using probes of different capacities (single-axis projection onto the difference between class means, logistic regression, and a small MLP), while setting up three image conditions as controls: original images without modification; `noise`, which replaces images with random noise to remove scene content while retaining sample diversity as a control for probe overfitting; and `gray`, which replaces images with the same uniform gray image to remove both scene content and differences between samples; across these conditions, we measure heading-decoding accuracy, the amplitude of image-dependent answer variation amp, and circular regression of azimuth (detailed definitions and results are in Appendix B). The results show that reference-object heading information is linearly decodable

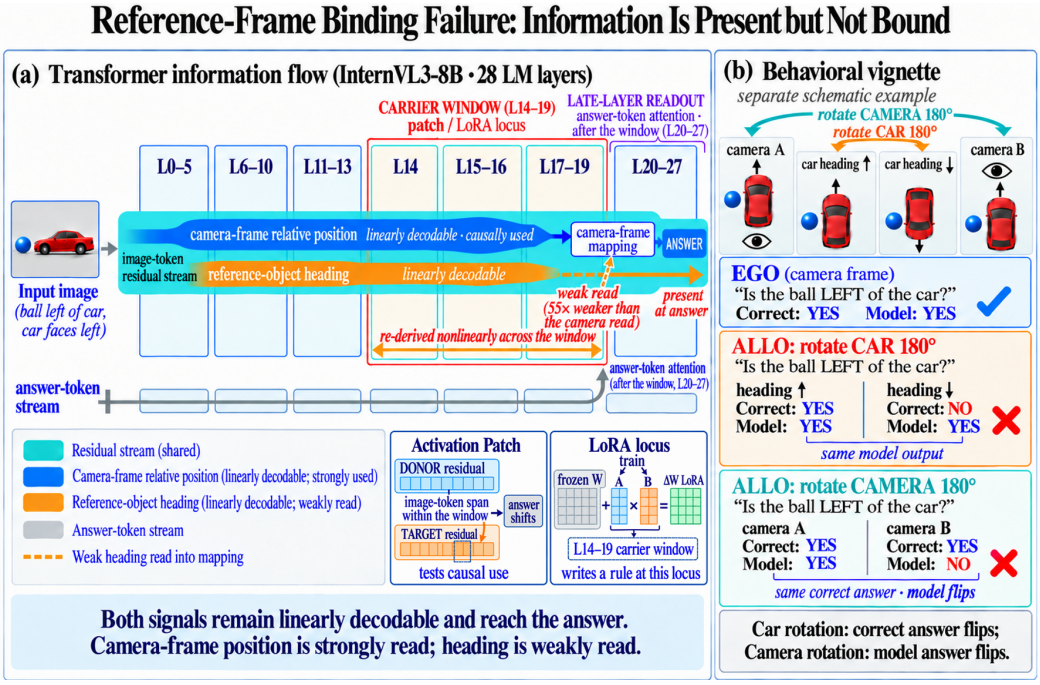

Figure 1: Mechanism chain and a separate behavioral schematic. Panel (a) uses InternVL3-8B's 28 language layers; the 55× label refers to its probability-scale camera/heading patch-effect ratio. Image-span interventions identify the L14–19 window. Heading remains decodable across this window, and later answer-token attention contributes to camera readout. The LoRA inset denotes $W' = W + \Delta W$, with a trained low-rank update $\Delta W = AB$. Panel (b) illustrates the different consequences of rotating the reference object and rotating the camera.

from the models' representations, so task failures cannot be explained simply by the absence of heading-related visual information.

We additionally introduce a phase-scrambled condition that disrupts image spatial structure while preserving only per-channel color histograms and the amplitude spectrum to further investigate how heading information is represented in the residual stream, and the results suggest that the residual stream carries a signal highly correlated with reference-object heading, with part of this signal carried by low-level color and spectral statistics rather than entirely by spatial structure.

5.2 Heading Information Is Transmitted

For the transmission stage, we decode heading information from the residual stream at the token position that predicts the answer using probes of different capacities (single-axis projection onto the difference between class means and L_2-regularized logistic regression), while varying regularization strength and the number of leading principal components available to logistic regression, and adding shuffled-label controls to test whether the probes' high decoding accuracy indeed depends on the correspondence between residual features and true heading labels (detailed settings and results are in Appendices C and L.5). Full-dimensional logistic regression achieves 0.9830–1.0000 accuracy, whereas shuffled-label controls remain near chance and restricting decoding to a single axis or a few leading principal components generally reduces accuracy, showing that heading-related information reaches the answer position and remains linearly decodable, and is represented in a high-dimensional subspace.

We additionally evaluate cross-object generalization by training probes on nine reference-object categories and testing on the remaining category, and find that heading-decoding accuracy falls to 0.481–0.537, while camera-relative target-position decoding remains at 0.844–0.947 (Table 16), indicating that the transmitted heading information does not generalize reliably to unseen reference objects under the tested linear probes.

5.3 USE OF HEADING INFORMATION

For the information-use stage, we separately reverse the components of the answer-position residual stream corresponding to heading information and the model's answer, and compare their effects (Li et al., 2023; Tan et al., 2024). With residual h and unit class-mean heading direction $\hat{w}_{\text{head}}$, the intervention is

$$h' = h - 2(\hat{w}_{\text{head}}^{\top} h)\hat{w}_{\text{head}}. \tag{2}$$

Writing answer margin as $m = \text{logit}_{\text{yes}} - \text{logit}_{\text{no}}$, we define four behavioral measures:

$$\begin{aligned} r_{\text{nat}} &= \frac{|\Delta m(\hat{w}_{\text{head}}, M_{\text{head}})|}{|\Delta m(\hat{w}_{\text{ans}}, M_{\text{ans}})|}, \qquad \Delta_{\text{beh}} = |\Pr(\hat{a} = \text{yes} \mid R) - \Pr(\hat{a} = \text{yes} \mid L)|, \\ d &= \frac{|\bar{m}_L - \bar{m}_R|}{s_{\text{within}}}, \qquad \text{SF} = \frac{1}{P}\sum_{j=1}^{P} \mathbf{1}[\hat{a}_j^L \neq \hat{a}_j^R]. \end{aligned} \tag{3}$$

Here $\Delta m(\hat{v}, M)$ is the answer-margin change when reversing the component along unit direction $\hat{v}$ at its observed magnitude M, and $\hat{w}_{\text{ans}}$ is obtained from the difference between residual class means grouped by the model's own answers. The ratio r_{nat} measures the effect of reversing the heading component relative to reversing the answer component. The other three measures compare responses to the same allocentric question on P heading-reversal image pairs: Δ_{beh} measures the difference in yes-answer proportions, d measures the difference in mean answer margin normalized by the within-condition standard deviation s_{within}, and SF measures the proportion of pairs with different discrete answers $\hat{a} \in \{\text{yes}, \text{no}\}$. Results from four different models show that heading information is strongly decodable at the answer position, yet the models make little use of it when answering (Appendix D; Tables 17 and 18).

We additionally measure the absolute cosines of the logistic-regression heading direction $\hat{w}_{\text{probe}}$ and the unit direction of answer information $\hat{w}_{\text{ans}}$ with $g = W_U[\text{yes}] - W_U[\text{no}]$, where W_U is the unembedding matrix and g the answer readout axis. Whether information along a latent direction enters the final answer depends on its component along g. We also measure

$$r_{\text{probe}} = \frac{|\cos(\hat{w}_{\text{probe}}, g)|}{|\cos(\hat{w}_{\text{ans}}, g)|}, \tag{4}$$

which measures the readout alignment of heading information relative to answer information. Results from eight models show that the heading direction is nearly orthogonal to the readout axis, and the four models with all three measures show much weaker alignment for the heading direction than for the empirical answer direction, indicating weak direct readout of the decodable heading information (Table 19).

Reference-object heading information is perceived by the models and transmitted to the position where the answer is generated, but is ultimately ignored rather than used; the task bottleneck lies in the final mapping from the latent vector to the answer.

5.4 COMPARING HEADING AND CAMERA INTERVENTIONS

To compare the use of heading and camera information, we replace a recipient image's image-token residual span with the corresponding donor span at the same layer and continue the forward pass, using heading-reversal and camera-reversal pairs under both allocentric and egocentric questions. Because egocentric answers are independent of reference heading, they provide a control for task-irrelevant heading changes. For pre-/post-patch yes probabilities p_i, p_i', we define effect size ϵ as

$$\epsilon = \frac{1}{N}\sum_i |p_i' - p_i|. \tag{5}$$

Here N is the number of evaluated inputs. We measure effect size ϵ and residual perturbation magnitude $\rho = \|\Delta h_{\text{img}}\|/\|h_{\text{img}}\|$ across conditions, measure its log-odds counterpart ϵ^{log}, and use a fine-tuned heading classifier to test measurement reliability (Appendix E).

The experimental results show that, when answering, the models rely strongly on camera-relative position in a causal sense and make little use of reference-object heading information, with answers more sensitive to camera-relative information. We additionally conduct information-erasure experiments to test whether residual perturbation magnitude affects these results (Appendix L).

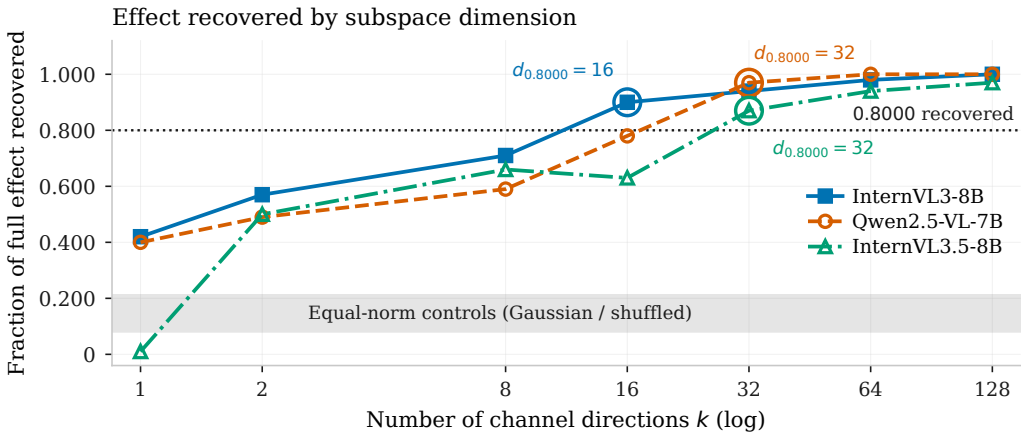

Figure 2: A low-dimensional camera carrier. Subspaces are fitted on training pairs and evaluated by five-fold cross-validation. Circles mark the 16–32 directions needed to recover 0.8000 of the full patch effect. Equal-norm Gaussian and channel-shuffled perturbations recover much less. Components are ordered by variance. Full effects and controls are reported in Appendix I.

5.5 Where camera information affects readout

To locate camera-information use, we apply camera-reversal patches at the visual output, language-model input, and successive image-token residuals, measuring ϵ at each site. In InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, visual-output effects are 0.591–0.743, matching language-model-input effects (Table 31). The layerwise effect declines over L14–19 in InternVL3-8B and Qwen2.5-VL-7B and L18–24 in InternVL3.5-8B; InternVL3-8B falls from 0.731 at layer 14 to 0.040 at layer 17 (Appendix H). These ranges identify where image-span replacement loses influence on the answer.

We then replace answer-position attention outputs with dataset means from the window start through progressively deeper layers. In the same three models, replacement through the window end suppresses camera sensitivity by 0.090–0.110, compared with 0.980–0.990 through the deepest layer (Table 30), placing most measured answer-position readout after the window. The same test shows earlier readout in llava-1.5-7B and idefics2-8B, with substantial suppression within their windows (Appendix H).

We further restrict patches to channel subspaces fitted from camera-reversal differences on training pairs, using five-fold cross-validation over 1,440 pairs and equal-norm Gaussian and channel-shuffled controls. In the three primary models, 16–32 directions recover 0.8000 of the full replacement effect (Figure 2), while full replacement is five to eight times stronger than the larger matched-norm control (Table 36), indicating that the effect depends on specific content concentrated in a small channel subspace.

To examine which components contribute, we compare individual and joint residual-update patches and cumulatively replace attention-head outputs with dataset means, using equally sized random head sets as controls. Within the readout window, individual update effects reach at most 0.063, whereas joint updates reach 0.243–0.429 (Table 32). The eight highest-contributing heads remove 0.380–0.590 of camera sensitivity, compared with 0.010–0.020 for random sets; even the top 32 leave substantial sensitivity, and nearly all heads in the measured band are needed for 0.970–0.990 suppression (Table 35). Thus, the compact camera representation is associated with distributed updates and readout across multiple heads.

5.6 Heading recovery after erasure

To test whether heading remains recoverable after removing its linearly decodable components, we fit heading subspaces on training samples and apply projection ablation before the readout window, comparing pooled and position-wise removal with matched random-subspace controls. Pooled

removal reduces input decoding accuracy to 0.605–0.687, and position-wise removal to 0.5388–0.6253, while output decoding returns to 1.000 in both tests (Appendix L.1). Because input decoding remains above chance, these results alone cannot distinguish recovery from residual linear information from nonlinear reconstruction.

We therefore apply LEACE using separately trained erasure operators, with random same-rank removal and camera-sector decoding as controls (Belrose et al., 2023). Window-entry heading decoding falls to 0.500, below shuffled-label thresholds, whereas random removal retains at least 0.9831 and heading erasure changes camera-sector decoding by at most 0.0068 (Table 47). In the subsequent recovery tests on InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, window-output decoding returns to 1.000 (Appendix L.3), showing that downstream processing restores linear access to heading from the remaining representation.

5.7 The answer mapping does not rotate with the object

For the answer-mapping stage, we measure whether the response pattern over target azimuth changes with reference heading, fitting answer margin separately for each model, spatial relation, and heading side as

$$m_c(\theta) = a_c \cos(\theta - \varphi_c) + b_c, \qquad c \in \{L, R\}. \tag{6}$$

Here θ is the target's azimuth relative to the camera, a_c is modulation amplitude, b_c is an offset, and φ_c is the peak-response azimuth. We measure the phase difference

$$\Delta\varphi = \mathrm{wrap}_{180}(\varphi_L - \varphi_R), \tag{7}$$

where wrapping maps angles into $[-180°, 180°]$. Correct object-centered binding predicts $|\Delta\varphi| = 180°$ under heading reversal, whereas a camera-centered response predicts $0°$. Because phase is unreliable at low amplitude, we require $\bar{a} = (a_L + a_R)/2 \geq 1.000$, using separately fitted amplitudes so opposite curves cannot cancel, and check the procedure with synthetic responses of known phase (Appendix F).

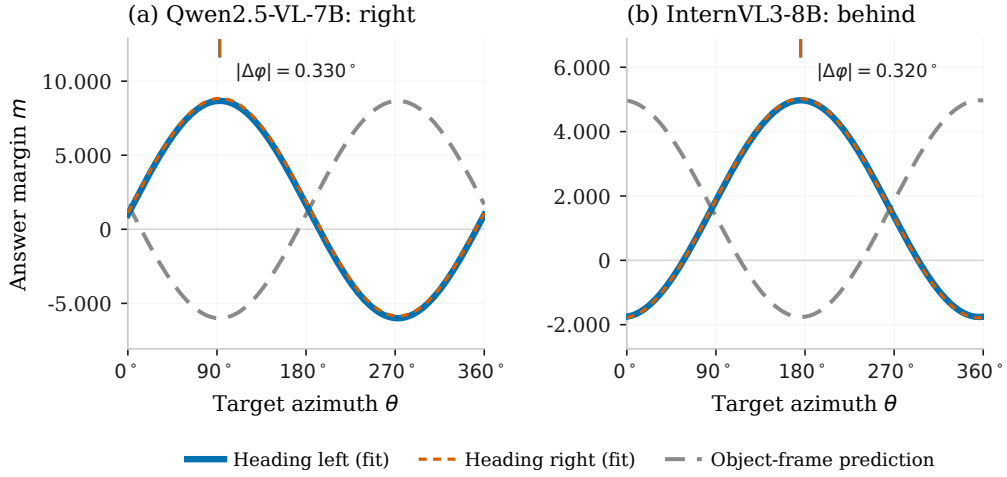

Figure 3: Separately fitted answer curves under opposite object headings. The Qwen2.5-VL-7B right relation and InternVL3-8B behind relation have phase differences of 0.330° and 0.320°. Dashed gray curves show the 180° rotation required by object-centered binding. These are fitted curves; Appendix F gives their parameters and all combinations.

Nineteen of 32 model–relation combinations pass the amplitude criterion, spanning seven models; their maximum phase difference is 1.990°, and raising the threshold to 2.000 or 3.000 leaves this maximum unchanged (Figure 3; Table 22). The synthetic camera-frame and correctly bound responses yield 0° and 180° without added noise, respectively. The measured response patterns therefore remain nearly fixed under heading reversal rather than undergoing the rotation required for binding.

We also assess fit quality and saturation to determine where phase is informative. The three primary models have median pooled-fit R^2 of 0.8109–0.9474, while fit is weaker in the others; Gemma-3-4B almost always answers yes, and no idefics2-8B combination passes the amplitude threshold (Appendix F). For these cases, accuracy and paired response changes assess whether nearly constant answers follow the changed labels, and phase alone is insufficient.

5.8 Fine-tuning transfers across objects but fails across headings

To test whether targeted training improves reference-frame binding, we fine-tune the localized window with reference-object categories separated between training and testing, comparing single-relation LoRA, four-relation LoRA, and full-parameter tuning. The objective combines allocentric answer classification, a penalty on incorrect paired probability differences, and preservation of egocentric predictions (Appendix G.3). All nine settings reach training $D \geq 0.832$, but test accuracy never exceeds 0.485; full tuning performs worst in every model, at no more than 0.111 (Table 1), showing that successful training fit does not yield transfer under this experimental setup.

Table 1: Fine-tuning with reference-object categories separated between training and testing. Each setting gives training directed correctness D and test accuracy. All nine runs use the three Qwen-family models. Full test D values and losses appear in Table 25.

	Single-relation LoRA		Four-relation LoRA		Full tuning	
Model	Train D	Test acc.	Train D	Test acc.	Train D	Test acc.
InternVL3-8B	1.000	0.361	0.958	0.293	1.000	0.056
InternVL3.5-8B	0.997	0.333	0.996	0.485	1.000	0.028
Qwen2.5-VL-7B	1.000	0.444	0.953	0.192	0.832	0.111

Because this category split requires both recognizing new objects and binding their headings, we additionally separate these demands by crossing seen/unseen objects with trained/reversed headings in four-choice direction questions. Training uses seven object categories with only left-facing references, with two InternVL3.5-8B seeds and one llava-1.5-7B seed, giving three checkpoints from two different language backbones. Each uses 420 questions, 15 epochs, rank-16 LoRA, and learning rate 0.0001, updating L18–24 in InternVL3.5-8B or L11–16 in llava-1.5-7B with direction-classification and egocentric-preservation losses. We measure accuracy and the antipodal rate p_{opp}, the proportion of predictions exactly $180°$ from the correct direction, whose random-choice baseline is 0.250.

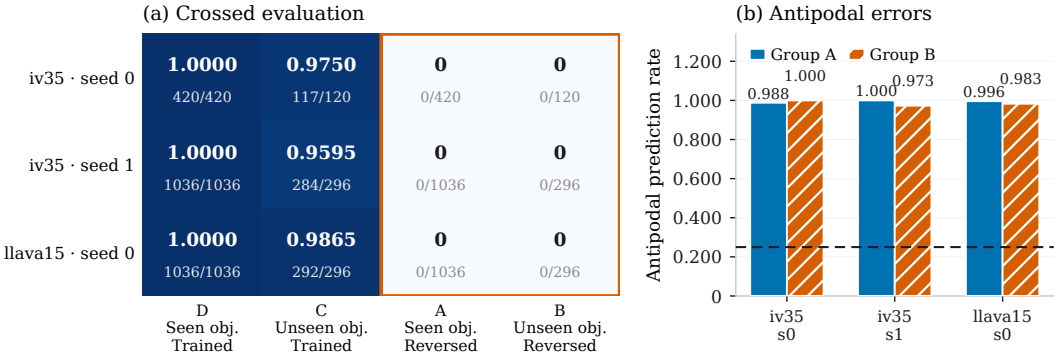

Figure 4: Crossed evaluation after fine-tuning. Unseen-object accuracy is 0.9595–0.9865 at the trained heading and zero at reversed heading. Reversed-heading errors are almost entirely antipodal; the dashed line marks chance (0.250). A/B denote seen/unseen objects at reversed heading. Three checkpoints cover two models. Full results appear in Appendix G.6.

At the trained heading, all three checkpoints reach 1.0000 accuracy on seen objects and 0.9595–0.9865 on unseen objects. Reversing heading reduces accuracy to zero in both groups, with 0.9730–1.0000 of predictions opposite the correct direction (Figure 4; Table 27). Since the unseen objects are

identical across the two heading conditions, these results separate successful object transfer from failure to adapt to reversed headings in the four-choice evaluation.

We further inspect the prediction distribution and performance at target azimuths absent from training. In the recorded InternVL3.5-8B seed-0 group, each direction is selected roughly a quarter of the time despite zero accuracy; for InternVL3.5-8B seed 1 and llava-1.5-7B seed 0, $22/37$ evaluated azimuths are unseen during training, yet seen-object accuracy remains 1.0000 (Appendix G.6). These results support a direction rule that transfers across objects and positions but continues using the heading from training.

5.9 Localization, text, reasoning, and erasure controls

To assess basic localization ability, we compare the models' judgments under egocentric and allocentric reference frames for the left and right spatial relations on SpinBench (Appendix J). Egocentric accuracy is 0.6993–0.9755, while allocentric accuracy remains near chance (Table 38). All tested models have higher egocentric than allocentric accuracy, so poor localization alone cannot explain these errors.

To further investigate whether other methods can help models answer correctly, we conduct a series of experiments (Appendix K). The results show that, whether the models describe the images themselves, receive the reference object's heading in text, or receive both explicit heading text and images, they fail to use heading information when answering allocentric questions; even explicitly instructing them to use it does not resolve the problem.

6 Limitations

Model coverage is limited to VLMs with at most 8B parameters. Larger models and native reasoning models remain to be tested. Real-photo experiments measure probe retraining, frozen-direction transfer, and steering using camera-relative pose annotations (Appendix M). Extending the causal analysis to photographs requires matched scenes with independently varied reference heading and camera position.

7 Conclusion

Across extensive experiments, the tested VLMs perform substantially better on egocentric questions, with some approaching perfect accuracy in basic localization, while allocentric judgments remain unreliable. Visual information about reference-object heading is encoded in their representations and transmitted to the answer position, yet has little influence on their answers, which fail to follow the object-centered frame specified by the instruction. In the detailed experiments, heading information is carried forward from earlier layers and written again in subsequent updates, and its linear decodability can recover downstream after its estimated linear components are erased, yet these processes do not yield reliable reference-frame binding. Camera-relative information is also carried forward and redundantly encoded across layers, then read out through multiple attention heads to strongly influence answers even under allocentric instructions; layerwise analyses in two models further show that the subspace through which this information affects answers changes during processing.

We further test alternative explanations across these stages through three groups of ablations and controls, examining whether the models possess the prerequisite abilities for the task (Appendix J), whether difficulties expressing answers or limited reasoning resources explain the failures (Appendix K), and whether our measurements and interventions are reliable (Appendix L). We investigate both internal changes through targeted fine-tuning and external support through descriptions, explicit heading statements, and additional reasoning. Fine-tuning transfers across objects but fails under heading reversal, while the tested prompting and reasoning settings do not reliably restore binding. Finally, we consolidate the experimental workflow and evaluation metrics in Appendix O to help future work assess whether proposed improvements establish reference-frame binding across objects and headings.

Reproducibility Statement

All experimental seeds and decision criteria will be released with the code. The appendices report model coverage, data splits, intervention definitions, control conditions, and full numerical results.

Generative AI Use Statement

Generative AI tools were used to polish the writing and to assist literature discovery. All AI-assisted text was checked against the reported results, and every AI-suggested reference was verified against the original source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Laura A. Carlson-Radvansky and Gordon D. Logan. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 1997.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *International Conference on Machine Learning (ICML)*, 2025.

Eve Danziger. Deixis, gesture, and cognition in spatial frame of reference typology. *Studies in Language*, 2010.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. GRIT: Teaching MLLMs to think with images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Andrew U. Frank. Formal models for cognition — taxonomy of spatial location description and frames of reference. In *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, volume 1404 of *Lecture Notes in Computer Science*, pp. 293–312. Springer, 1998.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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)*, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Stephen C. Levinson. Frames of reference and Molyneux's question: Crosslinguistic evidence. In *Language and Space*, pp. 109–170. MIT Press, 1996.

Stephen C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, 2003.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. doi: 10.48550/arXiv.2505.21500. URL https://arxiv.org/abs/2505.21500.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuanyuan Liu, Haiyang Mei, Dongyang Zhan, Jiayue Zhao, Dongsheng Zhou, Bo Dong, and Xin Yang. View-on-graph: Zero-shot 3D visual grounding via vision-language reasoning on scene graphs. In *AAAI Conference on Artificial Intelligence*, 2026.

Gordon D. Logan and Daniel D. Sadler. A computational analysis of the apprehension of spatial relations. *Language and Space*, 1996.

Liqin Luo, Guangyao Chen, Xiawu Zheng, Yongxing Dai, Yixiong Zou, and Yonghong Tian. Connecting the dots: Training-free visual grounding via agentic reasoning. In *AAAI Conference on Artificial Intelligence*, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan L. Yuille, and Jieneng Chen. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

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

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krähenbühl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations (ICLR)*, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zihan Weng, Lucas Gomez, Taylor Whittington Webb, and Pouya Bashivan. Caption this, reason that: VLMs caught in the middle. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhen Yang, Ping Jian, Zhongbin Guo, Zuming Zhang, Chengzhi Li, Yonghong Deng, Xinyue Zhang, and Wenpeng Lu. How do LLMs and VLMs understand viewpoint rotation without vision? an interpretability study. *arXiv preprint arXiv:2604.15294*, 2026.

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.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025a.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *International Conference on Learning Representations (ICLR)*, 2025b.

APPENDIX CONTENTS

Section	Experiments and contents	Page
A	**Behavioral evaluation**. Compare answers before and after reference-heading reversal and test preference for egocentric answers on SpinBench.	14
B	**Perception**. Decode heading from original, noise, gray, and spatially scrambled images; test how repeated images affect probe evaluation.	15
C	**Transmission**. Decode heading and spatial answers at the answer position; vary probe capacity and test on unseen objects.	20
D	**Information use**. Reverse heading and answer components of the residual stream; measure answer changes and alignment with the readout axis.	23
E	**Camera information**. Replace image activations using heading-reversal or camera-reversal pairs under both reference-frame instructions.	25
F	**Answer mapping**. Fit answer curves over target azimuth, compare their phases, and check the fitting procedure using curves with known phases.	27
G	**Targeted fine-tuning**. Separate generalization to unseen objects from generalization to reversed headings; compare training layers and loss functions.	31
H	**Layers that influence the answer**. Replace activations at successive layers and compare the resulting answer changes with attention to image tokens.	35
I	**How camera information affects answers**. Replace visual and language-model activations, separately and together, and test learned subspaces.	37
J	**Prerequisite abilities**. Test reference-frame instructions, object localization, and depth, including generalization across wording and images.	41
K	**Descriptions and reasoning**. Test model-generated descriptions, supplied heading statements, instruction following, and additional reasoning tokens.	45
L	**Measurement controls**. Remove heading or camera components and remeasure decoding and answers; test probe capacity and interventions with known effects.	47
M	**Real photographs**. Fit heading decoders on photographs, apply directions learned from rendered images, and measure their effects on spatial answers.	53
N	**Extended related work**. Spatial reference frames and prior work on spatial reasoning in vision-language models.	55
O	**Evaluation metrics**. Choose questions, image pairs, and measurements to assess binding, generalization, information use, and proposed improvements.	57

A Full Behavioral Methods and Results

We evaluate whether the models change their spatial answers when the question requires a different reference frame. COMFORT provides image pairs in which the reference object reverses its heading while the target position and camera view remain unchanged. SpinBench provides an independent set of questions that explicitly request either an observer-centered, egocentric judgment or an object-centered, allocentric judgment.

A.1 Models

The COMFORT heading-reversal experiment uses InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, and Gemma-3-4B. SpinBench adds LLaVA-OneVision-7B, llava-1.5-7B, Phi-3.5-Vision, and idefics2-8B, for eight models in total.

A.2 Heading reversal on COMFORT

For each COMFORT image pair, we ask the same allocentric yes/no question about the target relative to the reference object. The two images differ only in the reference object's heading, and we retain pairs for which the correct answer is yes in one image and no in the other. Let L denote the image with the reference object facing left and R the image with it facing right. Let $A_L^{(i)}$ and $A_R^{(i)}$ equal one when the corresponding answer is correct and zero otherwise. If the model gives the same discrete answer on both images, it must answer exactly one correctly, so

$$A_L^{(i)} + A_R^{(i)} = 1, \qquad \mathrm{acc}_{\mathrm{pair}} = \frac{1}{2P}\sum_i (A_L^{(i)} + A_R^{(i)}) = \frac{1}{2}. \tag{8}$$

Here P is the number of image pairs. We also calculate directed correctness D and sign agreement (SA), defined in Eq. (1), from the two yes probabilities and their correct labels. SA measures how often the probability changes in the direction required by the labels, while D measures the average signed size of that change.

Allocentric accuracy is close to 0.500, and D remains close to zero in all four models (Table 2). The probability changes provide little correctly directed response to heading reversal in this test.

Table 2: COMFORT responses to the same allocentric yes/no question on two images that differ only in reference heading and have opposite correct answers. SA is the fraction of pairs with a correctly directed probability change, accuracy scores individual answers, and D is the average signed probability change toward the correct answers. The last two entries show reference values for symmetric random predictions and probabilities equal to the correct binary labels.

Model	SA	Allocentric accuracy	D
InternVL3-8B	0.407	0.491	-0.001
InternVL3.5-8B	0.472	0.495	$+0.002$
Qwen2.5-VL-7B	0.470	0.496	0.000
Gemma-3-4B	0.500	0.504	$+0.004$
Symmetric random predictions (expectation)	0.500	0.500	0.000
Ideal predictions ($p = t$)	1.000	1.000	1.000

A.3 SpinBench evaluation

We use the single-image subset of SpinBench (Zhang et al., 2026): 926 multiple-choice questions, comprising 636 egocentric and 290 allocentric questions. For each image and question, the model's highest-scoring answer option is its prediction. We calculate accuracy separately for the two reference frames. To examine whether an incorrect allocentric answer instead follows the observer's perspective, we additionally use the 196 allocentric questions for which the corresponding egocentric answer can be identified unambiguously.

Table 3: SpinBench accuracy on 636 egocentric and 290 allocentric multiple-choice questions. Gap is egocentric minus allocentric accuracy. The final column computes this difference using only left/right questions that compare the two reference frames, reducing the contribution of depth-perception errors. For Gemma-3-4B, marked with an asterisk, interpretation uses this left/right comparison because its egocentric depth judgments are also weak; its overall gap is not reported.

Model	Egocentric	Allocentric	Gap	Left/right gap
InternVL3-8B	0.939	0.538	+0.401	+0.419
InternVL3.5-8B	0.942	0.572	+0.369	+0.401
Qwen2.5-VL-7B	0.934	0.528	+0.406	+0.380
LLaVA-OneVision-7B	0.923	0.486	+0.437	+0.386
Phi-3.5-Vision	0.810	0.455	+0.355	+0.393
idefics2-8B	0.695	0.528	+0.167	+0.156
llava-1.5-7B	0.689	0.438	+0.251	+0.261
Gemma-3-4B*	0.687	0.476	—	**+0.247**

Because errors in depth perception can lower egocentric accuracy as well as allocentric accuracy, especially for Gemma-3-4B, we additionally compare the two reference frames on left/right questions. These questions require the same type of object localization but differ in whether the answer is judged from the observer's perspective or the reference object's perspective. The left/right accuracy gap reduces the influence of front/back depth judgments on this comparison.

Define $\mathrm{gap} = \mathrm{acc}_{\mathrm{ego}} - \mathrm{acc}_{\mathrm{allo}}$. This measures the advantage of camera-relative over object-relative answering. The *egocentric substitution rate among allocentric errors* is

$$s_{\mathrm{ego}} = \frac{|\mathcal{S}|}{|\mathcal{W}|}, \qquad \mathcal{W} = \{\, i : \hat{y}_i \neq y_i^{\mathrm{allo}} \,\}, \qquad \mathcal{S} = \{\, i \in \mathcal{W} : \hat{y}_i = y_i^{\mathrm{ego}} \,\}, \tag{9}$$

Here $\hat{y}_i$ is the model's answer to an allocentric question. The labels y_i^{allo} and y_i^{ego} are the correct answers from the reference object's perspective and the observer's perspective, respectively. Within the 196-question subset, $\mathcal{W}$ contains the questions answered incorrectly under the requested allocentric frame, and $\mathcal{S}$ contains those errors that would be correct under the egocentric frame.

Egocentric accuracy exceeds allocentric accuracy in all eight models (Table 3). On the questions for which both reference-frame answers are identifiable, more than half of each model's allocentric errors match the egocentric answer (Table 4). These errors show a systematic preference for the observer's perspective.

Table 4: Egocentric substitution rate s_{ego} among allocentric errors, using Eq. (9). The analysis uses 196 SpinBench allocentric questions for which the correct egocentric answer can also be identified. Each value is the fraction of a model's incorrect allocentric answers that equal that egocentric answer.

Model	Substitution rate
InternVL3.5-8B	0.659
Qwen2.5-VL-7B	0.633
Phi-3.5-Vision	0.628
idefics2-8B	0.622
Gemma-3-4B	0.615
LLaVA-OneVision-7B	0.580
InternVL3-8B	0.531
llava-1.5-7B	0.524

B Full Perception Methods and Results

Heading decoding at image-token positions provides a test of what visual information is available before answer generation. We compare original images with noise and uniform gray inputs, and use

phase randomization to separate spatial structure from color and frequency statistics. A separate evaluation checks whether repeated images across questions inflate probe accuracy.

B.1 MODELS

The image-replacement and image-structure experiments use InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B. The repeated-image check also includes the other five models in Appendix C.1. It reuses image-token residuals saved by earlier feature-extraction runs for 57 evaluated model–object-configuration cases; residuals had not been exported for the remaining cases in the full model-by-configuration grid.

B.2 IMAGE-REPLACEMENT CONTROLS

For one COMFORT object configuration, we retain the original spatial questions and compare three image inputs. The `real` condition uses the original rendered image. The `noise` condition replaces each input image with independently generated random noise, removing scene content while keeping the inputs different from one another. The `gray` condition replaces every image with the same uniform gray image, removing both scene content and differences between images. Each condition contains 296 image–question samples, from 37 target azimuths, four queried spatial relations, and two reference headings.

To decode heading, we average the residual vectors over the image-token positions at a layer and train three probes on these vectors. The difference-of-means (DoM) probe projects each vector onto the difference between the mean vectors of the two heading classes. L_2-regularized logistic regression learns a linear heading classifier, and a small multilayer perceptron (MLP) also permits a nonlinear classifier. Probe accuracy measures how well the external decoder distinguishes left-facing from right-facing reference objects.

For the behavioral measurement, we ask the same egocentric yes/no question while changing the image. Within each group with fixed reference heading and queried relation, the 37 samples vary in target azimuth. We measure the range of the model's yes probability across these samples,

$$\mathrm{amp} = \max_i p_{\mathrm{yes}}^{(i)} - \min_i p_{\mathrm{yes}}^{(i)}, \tag{10}$$

and separately train a linear regression probe to predict the cosine and sine of target azimuth from the residual features,

$$R^2_{\mathrm{vec}} = 1 - \frac{\sum_i \|\hat{y}_i - y_i\|^2}{\sum_i \|\bar{y} - y_i\|^2}, \qquad y_i = (\cos\theta_i, \sin\theta_i). \tag{11}$$

Here i indexes images within the evaluated group, $p_{\mathrm{yes}}^{(i)}$ is the yes probability for image i, θ_i is the target's azimuth relative to the camera, $\hat{y}_i$ is the predicted two-component vector, and $\bar{y}$ is the mean of the true vectors. The quantity amp measures how much the model's answer probability varies as the target changes position. The subscript vec denotes the two-component target vector. The quantity R^2_{vec} measures how well target azimuth can be linearly decoded, compared with always predicting the mean vector. A score of zero matches that baseline, and a negative score performs worse.

Because perfect decoding on original images could also reflect information in prompts, position codes, or other variables correlated with the labels, we keep these inputs unchanged when replacing the image. A signal that depends on scene content should weaken after replacement, whereas a signal supplied entirely by the unchanged text should remain.

Because identical image inputs can produce identical residual features, we also count the number n_{distinct} of distinct feature vectors before fitting the probes. This separates removal of image content from the additional loss of sample variation caused by a uniform gray image.

All three probes use the same residual vectors averaged over image-token positions, and each reported accuracy is taken at that probe's best evaluated layer. The comparison between the two linear probes and the MLP tests whether a nonlinear decoder can recover information that the linear decoders miss.

For heading decoding in this image-replacement experiment, we use repeated random training/test splits and compare true labels with randomly reassigned labels under the same splits. These original random splits can place samples from a repeated image in both sets, so Appendix B.4 separately

checks accuracy after keeping all samples from an image together or retaining only one sample per image. The direct behavioral measurement amp requires no probe training. Circular regression is repeated with grouped training/test splits over $n_{\mathrm{seeds}} = 10$ random seeds; these image-replacement conditions share training/test divisions that permit some rendered images to occur in both sets.

Random-noise inputs retain 296 distinct residual vectors per model, whereas gray inputs collapse to only one or three vectors (Table 5); noise therefore provides the main comparison after scene content is removed. All nine model–probe combinations decode heading with accuracy 1.0000 on original images, but decoding falls near chance with noise or reassigned labels (Table 6). Gray inputs also yield near-chance decoding, with differently labeled samples sharing the same vector. Original images produce much larger answer-probability variation than either replacement; gray inputs give zero variation because all 37 images and the egocentric question are identical within each of the eight groups per model (Table 7). Target-azimuth regression remains high on original images and becomes negative on noise images in all three models (Table 8).

Table 5: Construction of the three image conditions and the number n_{distinct} of distinct image-token residual vectors among 296 samples. Counts are listed in the order InternVL3-8B / InternVL3.5-8B / Qwen2.5-VL-7B. Random noise removes scene content while retaining sample variation; the identical gray input removes both.

Condition	Construction	n_{distinct}
`real`	Original renders	216 / 216 / 72
`noise`	Independent noise per image	296 / 296 / 296
`gray`	Identical gray image	3 / 3 / 1

Table 6: Test accuracy for decoding left-facing versus right-facing reference objects from residual vectors averaged over image-token positions. Slash-separated values give the difference-of-means projection, L_2-regularized logistic regression, and multilayer perceptron, respectively, using each probe's best evaluated layer. The gray condition contains only one or three distinct vectors. The last entry gives the accuracy range after labels for the original-image features are randomly reassigned.

Model	`real`	`noise`	`gray`
InternVL3-8B	1.0000 / 1.0000 / 1.0000	0.5172 / 0.5169 / 0.5269	0.4561
InternVL3.5-8B	1.0000 / 1.0000 / 1.0000	0.5407 / 0.5777 / 0.5844	0.4561
Qwen2.5-VL-7B	1.0000 / 1.0000 / 1.0000	0.5642 / 0.5844 / 0.5677	0.4932
Shuffled labels (`real`)	0.4797–0.5406		

Table 7: Range amp of the yes probability when target azimuth varies across 37 inputs to the same egocentric question. Each model is evaluated in eight groups, with reference heading and spatial relation fixed within a group. The three columns use original images, an identical gray image, and independent random-noise images. All eight gray-image groups have zero range.

Model	`real`	`gray`	`noise`
InternVL3-8B	0.9849	0.0000	0.0976
InternVL3.5-8B	0.9998	0.0000	0.1189
Qwen2.5-VL-7B	0.9964	0.0000	0.0541

B.3 Phase-scrambling control

We transform each COMFORT image by randomizing the phase of its Fourier transform while retaining its amplitude spectrum, then match the histogram of each color channel to that of the original image. The phase changes obey Hermitian symmetry, meaning that paired frequency coefficients

Table 8: Linear decoding of the target's azimuth relative to the camera, scored by R^2_{vec} in Eq. (11). Each interval gives the minimum and maximum across $n_{\text{seeds}} = 10$ random seeds. Original-image and noise-image conditions use the same type of grouped training/test evaluation.

Model	`real`	`noise`
InternVL3-8B	$[0.9990,\ 1.0000]$	$[-1.0840,\ -0.9570]$
InternVL3.5-8B	$[0.9980,\ 1.0000]$	$[-0.9600,\ -0.8330]$
Qwen2.5-VL-7B	$[0.9840,\ 1.0000]$	$[-0.7670,\ -0.6950]$

remain complex conjugates so that the inverse transform produces a real-valued image. The operation disrupts where visual structures appear while preserving the specified color and frequency statistics. Unlike replacing every image with gray, this condition retains 296 distinct residual feature vectors. Probes use training/test groups defined by target azimuth, with all samples at the same azimuth assigned to the same group.

We first decode the target's azimuth from these features to check whether the transformation removed the spatial information captured by the circular-regression probe. We then decode the reference object's heading and compare its accuracy with the original-image and random-noise conditions. The questions and heading labels are unchanged when the image is transformed.

To quantify how much of the accuracy improvement from noise images to original images remains after phase randomization, we define

$$\text{ret} = \frac{A_{\text{scr}} - A_{\text{noise}}}{A_{\text{real}} - A_{\text{noise}}}, \tag{12}$$

Here $A_\bullet$ is binary heading-decoding accuracy. The subscripts scr, noise, and real refer to images after phase randomization, random-noise images, and original images, respectively. A value of zero means that the transformed-image accuracy equals the noise baseline; a value of one means that it equals the original-image accuracy.

Table 9: Decoding after randomizing Fourier phase while retaining amplitude spectra and matching color histograms. The first two columns report circular regression of target azimuth. The next three give binary heading-decoding accuracy for transformed, original, and noise images. Retention is ret in Eq. (12), the fraction of the accuracy improvement from noise to original images that remains after the transformation.

Model	R^2_{vec}		Heading accuracy			Retention ret
	Scrambled	Real	Scrambled	Real	Noise	
InternVL3-8B	-0.2241	$+0.9866$	0.7380	1.0000	0.5500	0.4200
InternVL3.5-8B	-0.1362	$+0.9764$	0.6750	1.0000	0.5500	0.2800
Qwen2.5-VL-7B	-0.2392	$+0.9777$	**0.8790**	1.0000	0.5500	**0.7300**

Phase randomization makes target-azimuth regression negative, while heading decoding remains above the noise baseline in all three models (Table 9). Part of the heading signal survives the loss of spatial structure; global color and frequency statistics likely carry it.

B.4 Checking whether repeated images inflate probe accuracy

A rendered image can be used with more than one spatial question. If samples from the same image appear in both the probe's training and test sets, a probe may recognize an image it has already seen instead of learning a heading distinction that generalizes to new images. The original training/test split options assigned individual samples randomly, grouped them by the question-relative angle index, or placed all questions about one relation in the test set. The angle index is the target's position in the 37-step scan stored separately for each queried relation: every scan runs from -180° to 180°, but its angular origin changes with the relation. For the same rendered target position, the stored

angles for front, left, and right questions differ from the behind-question angle by $-180°$, $+90°$, and $-90°$, respectively, modulo $360°$. None of these rules guarantees that all copies of an image stay together, because the same image can have different angle indices or relation labels in different question records.

Randomly reassigning a heading label to each image–question sample can give different labels to identical image features. A probe can then fall to chance on the randomly reassigned labels even though its high accuracy with the original labels came from remembering repeated images.

We compare four ways to construct training and test sets, keeping features, probes, cross-validation fold counts, and dimensionality-reduction settings fixed: (i) assign individual samples randomly; (ii) group samples by the original question-relative angle index; (iii) identify each rendered image from reference heading and target azimuth, then assign all its samples to the same set; and (iv) retain one question sample per image before dividing the distinct images into training and test sets. In the repeated-image configuration, the fourth procedure reduces 296 samples to 72.

To check that this comparison detects accuracy obtained by remembering repeated inputs, we additionally construct a dataset of random feature vectors. Each image identity receives a random vector and a label, and repeated samples of that identity share the same vector and label. There is no image content from which to infer heading, and the vectors provide no systematic heading rule for unseen identities. A probe can nevertheless remember the vector–label associations when copies of a vector appear in both training and test sets. We apply the same splitting methods to these data. The difference between splitting methods reflects both repeated-input overlap and the change in training/test composition.

For random vectors with repeated identities, accuracy falls from 1.0000 under the original angle-index split to 0.6273 when all copies are kept together and 0.5643 when only one copy is retained (Table 10). For the real image-token features, all 57 evaluated model–object combinations remain at 1.0000 under both procedures. The comparison detects the benefit of repetition in the constructed dataset, while removing repeated-image overlap does not reduce the real-data scores.

Table 10: Heading-decoding accuracy under four training/test divisions. Random assigns individual samples randomly; angle index uses the original question-relative index; image grouped keeps every sample from the same rendered image together; one per image retains only one sample per image before splitting. The first two entries use real model residuals averaged over image tokens and at the answer position, respectively. The final entry uses repeated random vectors whose labels can be remembered only when identities overlap between training and test sets. Keeping identities together reduces its accuracy by 0.3727. The random-split value for these vectors was not reported.

Data / readout site	Random	Angle index	**Image grouped**	**One per image**
Real image span (57 valid combinations)	1.0000	1.0000	**1.0000**	**1.0000**
Answer position (3 models)	1.0000	1.0000	**1.0000**	**1.0000**
Random vectors with repeated identities	n/a	1.0000	**0.6273**	**0.5643**

The first eight evaluated model–object combinations all retain accuracy 1.0000 after keeping image identities separate or retaining only one sample per image (Table 11). The distinct-vector count is the number of different residual feature vectors among the 296 image–question samples.

We also hold reference heading and target azimuth fixed and compare the residual vectors averaged over image tokens when the question asks about each of the four spatial relations. The relative distance between any two vectors in the same image group is exactly zero for Gemma-3-4B, llava-1.5-7B, idefics2-8B, and LLaVA-OneVision-7B. In these four models, changing only the queried relation does not change the measured image-token representation.

For InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, all four splitting methods give heading-decoding accuracy 1.0000 on the available features at the final sequence-token position (Table 10).

Table 11: Heading-decoding accuracy for the first eight evaluated model–object combinations under the four divisions defined in Table 10. Unique vectors counts distinct residual feature vectors among 296 samples. All four divisions give accuracy 1.0000, so the accuracy difference is 0 for every listed combination.

Model	Object	Unique vectors	Random	Angle index	Image grouped	One per image
InternVL3-8B	sofa	296	1.0000	1.0000	1.0000	1.0000
InternVL3.5-8B	sofa	296	1.0000	1.0000	1.0000	1.0000
Qwen2.5-VL-7B	sofa	296	1.0000	1.0000	1.0000	1.0000
Phi-3.5-Vision	bicycle	216	1.0000	1.0000	1.0000	1.0000
Gemma-3-4B	bicycle	**72**	1.0000	1.0000	1.0000	1.0000
llava-1.5-7B	bicycle	**72**	1.0000	1.0000	1.0000	1.0000
idefics2-8B	bicycle	**72**	1.0000	1.0000	1.0000	1.0000
LLaVA-OneVision-7B	bicycle	**72**	1.0000	1.0000	1.0000	1.0000

Preventing samples from the same image from appearing in both training and testing leaves these answer-position scores unchanged as well.

C Full Transmission Results

We run the model on a COMFORT image and spatial question, then extract the residual vector at the token position whose output predicts the answer. We refer to this vector as the **answer-position representation**. We train external probes on these vectors to predict either reference heading or the correct allocentric answer. The first test asks whether heading-correlated information remains available at the answer position; the second asks whether the available features support decoding the answer that the requested reference frame requires.

C.1 Models

We evaluate eight models: InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, LLaVA-OneVision-7B, llava-1.5-7B, idefics2-8B, Phi-3.5-Vision, and Gemma-3-4B.

C.2 Heading decoding at the answer position

We use $n = 296$ COMFORT answer-position samples and train a binary probe to predict whether the reference object faces left or right. Because a probe with too little capacity can fail to decode information that is present, we compare probe capacities while holding the features and training/test groups fixed. Cross-validation assigns both opposite-heading versions of a scene to the same group, so that one version cannot be used to predict its partner in the test set. The residual vectors have 3,584 dimensions for InternVL3-8B, Qwen2.5-VL-7B, and LLaVA-OneVision-7B, and 4,096 for InternVL3.5-8B.

The full-dimensional probe is L_2-regularized logistic regression. We vary its inverse regularization parameter C from 10^{-3} to 10^{2}. Smaller C penalizes large classifier weights more strongly, while larger C permits a more flexible fit. The four-model sweep includes the reported $C = 1$ setting (Table 12).

We compare full-dimensional logistic regression with projection onto the difference between the mean feature vectors of the two heading classes, which gives DoM accuracy 0.5203–0.8920 across eight models (Table 13). We also apply principal component analysis (PCA), rank components by the variance they explain, and train logistic regression using only the first k components. For logistic regression, feature standardization and PCA, when used, are fitted only on each training fold and then applied to its test fold. The retained count is $\min(k, d_{\text{model}}, n_{\text{train}} - 1)$, where d_{model} is the residual width and n_{train} is the training-sample count; Full uses $\min(d_{\text{model}}, n_{\text{train}} - 1)$. All probe

Table 12: Accuracy of full-dimensional logistic regression for decoding reference heading from the answer-position residual features. C is inverse regularization strength; smaller values impose a stronger penalty on classifier weights. Training and test groups keep opposite-heading versions of a scene together. Controls that randomly reassign heading labels remain near chance at every setting. The highlighted setting is $C = 1$.

C	InternVL3-8B	InternVL3.5-8B	Qwen2.5-VL-7B	LLaVA-OneVision-7B
10^{-3}	0.8208	0.8921	0.8446	0.9324
10^{-2}	0.9595	0.9492	0.9697	0.9931
10^{-1}	0.9832	0.9695	1.0000	0.9966
10^{0}	**0.9866**	**0.9729**	**1.0000**	**1.0000**
10^{1}	0.9866	0.9830	1.0000	0.9966
10^{2}	0.9832	0.9491	0.9932	0.9829

scores are averaged across grouped test folds. Varying k tests whether heading is recoverable from the largest-variance directions or requires additional lower-variance directions.

Table 13: Heading-decoding accuracy at the answer position for eight models. The first two columns give full-dimensional logistic regression at $C = 10^{-3}$ and its best reported setting. DoM projects onto the difference between the two heading-class mean vectors. Full PCA retains $\min(d_{\text{model}}, n_{\text{train}} - 1)$ components in each training fold, where d_{model} is the residual width and n_{train} is the training-sample count. Low-k minimum is the lowest accuracy over the tested principal-component counts.

Model	$C = 10^{-3}$	Best logistic	DoM	Full PCA	Low-k minimum
InternVL3-8B	0.8208	0.9866	0.5203	0.9866	0.5033
InternVL3.5-8B	0.8921	0.9830	0.5503	0.9762	0.5068
Qwen2.5-VL-7B	0.8446	1.0000	0.5233	1.0000	0.5000
LLaVA-OneVision-7B	0.9324	1.0000	0.5508	1.0000	0.4966
llava-1.5-7B	1.0000	1.0000	0.6251	1.0000	0.5034
idefics2-8B	0.9864	1.0000	0.6689	1.0000	0.5170
Phi-3.5-Vision	0.9967	1.0000	0.7633	1.0000	0.4931
Gemma-3-4B	0.9933	1.0000	**0.8920**	1.0000	0.5068

Table 14: Heading-decoding accuracy after retaining the k leading principal components of the answer-position features and fitting logistic regression. Components are fitted to training samples and ordered by explained variance. Full retains $\min(d_{\text{model}}, n_{\text{train}} - 1)$ components, where d_{model} is the residual width and n_{train} is the training-sample count. We report four models from the same recorded sweep used for Tables 12 and 13. Randomly reassigning heading labels gives near-chance accuracy at every k.

k	InternVL3-8B	InternVL3.5-8B	Qwen2.5-VL-7B	LLaVA-OneVision-7B
2	0.5033	0.5101	0.5036	0.5001
4	0.5406	0.5203	0.5171	0.4966
8	0.5136	0.5068	0.5000	0.5100
16	0.5206	0.5684	0.5640	0.6016
32	0.6590	0.7541	0.8451	0.9728
64	0.9729	0.9424	0.9967	0.9966
Full	0.9866	0.9762	1.0000	1.0000

To test whether high accuracy depends on the relationship between residual features and true heading, we randomly reassign the heading labels 1,000 times and refit the probe on the same answer-position features. These repetitions form a distribution of scores obtained after breaking the feature–heading

correspondence. In all seven models with this test, accuracy with the true labels exceeds this comparison distribution ($p \leq 0.0033$).

For the four models in that sweep, using only a few leading components gives low accuracy, whereas retaining 32 or more components substantially improves decoding (Table 14). Full-dimensional logistic regression also outperforms the class-mean projection in Table 13. The heading information recoverable by these probes is not generally concentrated along the class-mean difference or the first few directions with the largest feature variance.

Gemma-3-4B has the highest DoM accuracy, 0.8920, in Table 13. This single class-mean direction supports stronger heading decoding in Gemma-3-4B than in the other tested models.

C.3 Decoding the correct allocentric answer

To test whether the answer-position features support combining heading and target position, we train probes to predict the correct yes/no label of an allocentric question from the answer-position residual. COMFORT supplies 296 samples per object configuration (148 heading-reversal pairs), for $N = 2960$ samples across ten configurations. For scoring, we use the subset of pairs whose correct allocentric answers differ between headings; answering these pairs correctly requires using heading.

We evaluate two kinds of generalization. In the first, every sample at a given target azimuth belongs to the same training/test group, including both heading versions and repeated uses of an image. This prevents a test image's opposite-heading partner from remaining in training. In the second, we train on nine reference-object categories and test on the remaining category, repeating the procedure for each category. The first tests decoding at target azimuths absent from training, while the second tests transfer to reference objects absent from training.

We compare L_2-regularized logistic regression with a small MLP. Because arbitrary label reassignment would destroy the relationship between the two members of a heading-reversal pair, the comparison baseline permutes labels within heading-reversal pairs instead. We repeat this construction and use the 95th percentile of its decoding scores as the baseline. The pooled experiment evaluates four layers and uses ten repetitions for the linear and baseline measurements and five for the MLP measurement. We record the highest accuracy across the evaluated layers and compare final-layer linear accuracy with the final-layer baseline.

Table 15: Probe accuracy for the correct allocentric answer at the answer position, using eight models and ten pooled object configurations ($N = 2960$). All samples at the same target azimuth are assigned to the same training/test group, and scoring uses pairs whose correct answers reverse with heading. Linear max. and MLP max. are maxima across four evaluated layers; Linear final uses the final layer. Pair-swap final is the final-layer comparison value obtained by exchanging labels within pairs, reported as the 95th percentile of repeated baseline scores.

Model	Linear max.	Linear final	Pair-swap final	MLP max.
InternVL3-8B	0.6160	0.5960	0.5120	0.7920
InternVL3.5-8B	0.6500	0.6500	0.5110	0.8900
Qwen2.5-VL-7B	0.6390	0.6140	0.5090	0.8350
llava-1.5-7B	0.7230	0.6610	0.5190	0.8870
LLaVA-OneVision-7B	0.7210	0.7080	0.5100	0.8790
idefics2-8B	0.7920	0.7300	0.5180	0.9100
Phi-3.5-Vision	0.7030	0.6580	0.5150	0.8850
Gemma-3-4B	0.5650	0.5370	0.5170	0.7710

The highest linear accuracy across the evaluated layers is 0.5650–0.7920, and the highest MLP accuracy is 0.7710–0.9100 (Table 15). These scores show that trained decoders recover information about the correct allocentric answer from answer-position features.

We train separate logistic-regression probes for reference heading, the model's own answer, the correct egocentric answer, target position relative to the camera, and the correct allocentric answer. Training on nine reference-object categories and testing on the remaining category reduces heading accuracy

to 0.4810–0.5370, while camera-relative target-position accuracy is 0.8440–0.9470 (Table 16). This probe does not recover a reliable heading rule that transfers across reference objects, even though it can recover other spatial information from the same answer-position features.

Table 16: Decoding with L_2-regularized logistic regression trained on nine reference-object categories and tested on the remaining category. Separate probes predict reference heading, the model's own yes/no answer, the correct egocentric answer, target position relative to the camera, and the correct allocentric answer. Model-answer decoding checks whether the features retain the model's own output, while the other columns test transfer of the corresponding heading or spatial distinction to new objects. Gemma-3-4B always predicts the same answer, so its model-answer score of 1.0000 comes from a constant answer class.

Model	Heading	Model answer	Egocentric answer	Camera-relative position	Allocentric answer
InternVL3-8B	0.5020	0.9960	0.9290	0.9320	0.5000
InternVL3.5-8B	0.4990	0.9860	0.9200	0.9150	0.5010
Qwen2.5-VL-7B	0.4870	0.9920	0.9350	0.9470	0.4990
LLaVA-OneVision-7B	0.5210	0.9860	0.9000	0.9010	0.5410
llava-1.5-7B	0.5210	0.9430	0.8600	0.9270	0.5220
idefics2-8B	0.5240	0.9700	0.8020	0.8440	0.5120
Phi-3.5-Vision	**0.4810**	0.9790	0.7880	0.8870	0.5020
Gemma-3-4B	**0.5370**	1.0000	0.7510	0.8960	0.5080

D Full Information-Use Methods and Results

Decoding heading from a residual vector does not establish that the model uses it to answer. On COMFORT allocentric yes/no questions, we compare reversals of selected answer-position residual components with ordinary responses to heading-reversal image pairs. We also measure alignment between the decoded heading direction and the yes/no output direction to connect these answer effects with the readout geometry.

D.1 Models

The ratio r_{nat}, yes-answer proportion difference Δ_{beh}, and standardized answer-margin difference d use InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, and LLaVA-OneVision-7B. The paired answer-change rate SF additionally includes Gemma-3-4B. Geometric analysis also includes llava-1.5-7B, idefics2-8B, and Phi-3.5-Vision, giving eight models in total.

D.2 Data and splits

One COMFORT object configuration provides 296 image–question samples from 37 target azimuths, four queried relations, and two reference headings. These form 148 pairs. Within each pair, the target position, camera view, and allocentric yes/no question are unchanged, while the reference object reverses heading. We retain pairs whose correct answers are opposite, because answering both correctly then requires a heading-dependent judgment. The symbols L and R denote left-facing and right-facing reference objects, respectively.

Because probabilities near zero or one can conceal changes in output scores, we additionally use answer margin $m = \mathrm{logit}_{\mathrm{yes}} - \mathrm{logit}_{\mathrm{no}}$. Positive m favors yes and negative m favors no. Probes used for the geometric analysis are trained and evaluated by grouping opposite-heading versions of the same scene together, so those versions do not appear on different sides of the training/test split.

D.3 Behavioral measures

For the internal intervention, we obtain the unit heading direction $\hat{w}_{\mathrm{head}}$ by normalizing the difference between residual means for the two heading classes, and the unit answer direction $\hat{w}_{\mathrm{ans}}$ by normalizing

the difference between residual means grouped by the model's own yes/no answers. We reverse the residual component along each direction separately using Eq. (2), then continue computation and measure the change in answer margin. The ratio r_{nat} in Eq. (3) divides the absolute change caused by reversing the heading component by the absolute change caused by reversing the empirical answer component. Each component is reversed at its observed magnitude. A small ratio means that reversing the heading component affects the answer much less than reversing the component associated with the model's own answer.

The other three measurements use ordinary model responses to heading-reversal image pairs, without editing any internal vector. With discrete prediction $\hat{a} \in \{\mathrm{yes}, \mathrm{no}\}$, answer margin m, and within-heading standard deviation s_{within}, we define

$$\Delta_{\mathrm{beh}} = |\Pr(\hat{a} = \mathrm{yes} \mid R) - \Pr(\hat{a} = \mathrm{yes} \mid L)|, \qquad d = \frac{|\bar{m}_L - \bar{m}_R|}{s_{\mathrm{within}}}, \tag{13}$$

$$\mathrm{SF} = \frac{1}{P} \sum_{j=1}^{P} \mathbf{1}[\hat{a}_j^L \neq \hat{a}_j^R]. \tag{14}$$

Here P is the number of image pairs. The probability in Δ_{beh} is the proportion of discrete yes answers among images with a given heading. The quantity d compares the mean answer margins for the two headings relative to their variation within a heading condition. The paired rate SF counts how often the model changes its discrete answer on the two images. These three quantities measure sensitivity to heading reversal, while D and SA measure whether the changes follow the correct labels.

Reversing the heading component has a small effect relative to reversing the answer component, and ordinary heading reversal produces small changes in yes-answer proportions and mean answer margin (Table 17).

Table 17: Three measures of heading use for allocentric yes/no questions, pooled across four spatial relations. The ratio r_{nat} compares answer-margin changes after separately reversing the heading and empirical answer components of the residual. The quantity Δ_{beh} is the difference in discrete yes-answer proportions between reference headings. The quantity d divides the absolute difference in mean answer margin by its within-heading standard deviation.

Model	r_{nat}	Δ_{beh}	d
InternVL3-8B	0.0160	0.0070	0.0090
InternVL3.5-8B	0.0240	0.0200	0.0120
Qwen2.5-VL-7B	0.0110	0.0140	0.0130
LLaVA-OneVision-7B	0.0350	0.0070	0.0420

For the paired answer-change rate, we evaluate each of the four relations separately in five models, giving 20 model–relation combinations. Each model's largest SF across these relations gives its greatest observed sensitivity (Table 18).

Table 18: Largest paired answer-change rate SF among the four spatial relations for each model. For a given relation, SF is the fraction of heading-reversal image pairs on which the model gives different discrete yes/no answers to the same allocentric question. These measurements use unmodified model responses.

Model	Maximum SF
InternVL3-8B	0.0810
InternVL3.5-8B	0.0540
Qwen2.5-VL-7B	0.0270
LLaVA-OneVision-7B	0.0540
Gemma-3-4B	0.0000

For Gemma-3-4B, answer margin remains between $+9$ and $+11$, corresponding to $p_{\text{yes}} \approx 0.9999$, with very little variation within a heading condition. Dividing by that small variation gives $d = 0.2600$ and a relation-specific maximum of 0.8000, even though the relative change in answer margin is only 0.0480, with maximum 0.0720, and the score stays far above the yes/no boundary at zero. Table 18 accordingly shows no discrete answer changes for this model. Because this small denominator makes d difficult to compare with the other models, Gemma-3-4B is omitted from Table 17.

We further compare Gemma-3-4B's answer-margin change on questions whose correct answer reverses with heading and questions whose correct answer stays unchanged. The mean changes have the same sign and similar sizes, with ratio 0.9500. The small movement is also present when the task does not require an answer change, consistent with sensitivity to the altered rendering rather than selective use of heading for the requested relation. We use the unchanged discrete answers, measured by SF, for the model comparison.

D.4 Geometric measures

The unembedding matrix W_U converts the final normalized residual vector into scores for output tokens. The difference between its yes and no weight vectors defines the direction

$$g = W_U[\text{yes}] - W_U[\text{no}] \tag{15}$$

We account for the final normalization when expressing this output direction in residual-vector coordinates. In that local approximation, a change along a residual direction affects the yes/no score difference through its projection onto g. We therefore compare g with the direction that best decodes heading and with the direction associated with the model's own answer.

We obtain the unit heading direction $\hat{w}_{\text{probe}}$ by normalizing the coefficient vector of an L_2-regularized logistic-regression probe fitted to final-layer answer-position features. The training/test groups keep opposite-heading versions of a scene together. Test accuracy is at least 0.9730 and falls to chance when heading labels are randomly reassigned, confirming that this fitted direction decodes heading in the evaluated samples. We obtain the unit answer direction $\hat{w}_{\text{ans}}$ by subtracting the mean residual for samples on which the model answers no from the mean residual for samples on which it answers yes, then normalizing this difference. With $\cos(u, v) = u^\top v/(\|u\|\|v\|)$, we measure the absolute cosine of each direction with g and the ratio

$$|\cos(\hat{w}_{\text{probe}},\, g)|, r_{\text{probe}} = \frac{|\cos(\hat{w}_{\text{probe}},\, g)|}{|\cos(\hat{w}_{\text{ans}},\, g)|}. \tag{16}$$

An absolute cosine of one means that the directions are parallel, while zero means that they are orthogonal in this geometric approximation. Because the empirical answer direction is not itself perfectly aligned with g, the ratio r_{probe} compares heading alignment with answer alignment. A small value means that the decoded heading direction has much less direct projection onto the yes/no output direction than the empirical answer direction does.

The heading-to-output absolute cosine is at most 0.0132 across eight models (Table 19). In the four models with both direction cosines, the heading cosine is much smaller than the empirical answer cosine, and r_{probe} remains small. These geometric comparisons indicate weak alignment between the decoded heading direction and the final yes/no readout axis.

Gemma-3-4B gives the same answer class to every evaluated allocentric question. One of its two answer-class groups is empty, so the difference between answer-class means and the ratio using that direction are undefined. We report only its heading-to-output cosine.

E Comparing the Effects of Camera Position and Reference Heading

Reference heading and camera position define two distinct changes to an image. We compare their effects on answers by replacing the corresponding internal representations, using the same replacement procedure under allocentric and egocentric questions.

Table 19: Absolute cosines with the yes/no output direction g and their ratio at the final answer position. The first column uses the fitted logistic-regression heading direction. The second uses the normalized difference between residual means grouped by the model's own yes/no answers. Their ratio is r_{probe} in Eq. (4). The first measure is available for eight models and all three measures for four models; blank cells indicate values not reported or, for Gemma-3-4B's answer direction, undefined.

Model	$\lvert\cos(\hat{w}_{\mathrm{probe}}, g)\rvert$	$\lvert\cos(\hat{w}_{\mathrm{ans}}, g)\rvert$	r_{probe}
InternVL3-8B	0.0061	0.1820	0.0340
InternVL3.5-8B	0.0001	0.1240	0.0010
Qwen2.5-VL-7B	0.0067	0.2290	0.0290
LLaVA-OneVision-7B	0.0132	0.1500	0.0880
llava-1.5-7B	0.0015		
idefics2-8B	0.0132		
Phi-3.5-Vision	0.0046		
Gemma-3-4B	0.0097		

E.1 Models

The primary models are InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B. Their image-token positions retain a correspondence with locations in the image, and their yes probabilities are not consistently close to zero or one. We use these models for the main comparison of activation replacements. The other five models are tested in supplementary comparisons to examine how far the findings extend beyond the primary models.

E.2 Data and intervention

Each input contains one image and one spatial question requiring a yes/no answer. For an image pair, the donor is the image from which activations are copied, and the recipient is the image whose forward pass is modified. We first run the donor with the question and save the residual vectors at all image-token positions in a specified layer. We then run the recipient with the same question, replace its vectors at those positions with the donor vectors, leave the other token positions unchanged, and continue the forward pass to the answer (Vig et al., 2020; Meng et al., 2022). We compare the recipient's yes probability before and after this replacement.

We test four conditions. Under allocentric questions, reversing the reference heading changes the correct answer, whereas moving the camera should preserve the relation in the reference object's coordinates. Under egocentric questions, reversing reference heading should preserve the correct answer, whereas moving the camera changes the relevant observer-centered coordinates. The allocentric heading condition tests use of task-relevant heading information. The egocentric heading condition measures changes caused by a heading intervention when heading is irrelevant to the requested answer. The allocentric camera condition tests dependence on camera-relative information despite the object-centered instruction, and the egocentric camera condition checks that the intervention can detect dependence on camera position when that information is relevant.

COMFORT supplies 148 *heading-reversal pairs*, in which the reference object faces opposite directions while the scene, camera view, and target position remain fixed. It also supplies 144 *camera-reversal pairs*, in which the camera azimuth changes by 180° while object headings and world positions remain fixed. Within each pair, the question wording and queried spatial relation are unchanged.

Let h_{img} contain the recipient residual vectors at the image-token positions, and let Δh_{img} be the donor vectors minus these recipient vectors at the same layer. The effect ϵ in Eq. (5) is the mean absolute change in the recipient's yes probability across image pairs. We also measure $\rho = \|\Delta h_{\mathrm{img}}\|/\|h_{\mathrm{img}}\|$ to compare how much each replacement changes the residual representation.

We add two checks on this measurement. First, we calculate the log-odds effect $\epsilon^{\log}$ by using $\log[p/(1-p)]$, in place of yes probability p; this checks whether probabilities near zero or one

compress an otherwise detectable change (Zhang & Nanda, 2024). Second, we apply heading replacement to a fine-tuned heading classifier whose responses depend on the reference heading while ignoring which spatial relation is asked. This provides a positive control for detecting the classifier's known dependence on heading.

E.3 Results for the question and image-change conditions

Table 20: Effects ϵ of activation replacement in three primary models. Allo./camera denotes allocentric questions with camera-reversal donors; Allo./heading denotes allocentric questions with heading-reversal donors; Ego./heading denotes egocentric questions with heading-reversal donors. Each value is the mean absolute change in yes probability across the corresponding image pairs. The final column divides the allocentric camera-reversal effect by the allocentric heading-reversal effect.

Model	Allo./camera	Allo./heading	Ego./heading	Ratio
InternVL3-8B	**0.767**	0.014	0.018	$\sim 55.000\times$
InternVL3.5-8B	**0.692**	0.031	0.041	$\sim 22.000\times$
Qwen2.5-VL-7B	**0.704**	0.023	0.019	$\sim 31.000\times$

Under allocentric questions, heading-reversal effects are 0.014–0.031, while camera-reversal effects reach 0.692–0.767 (Table 20). Heading effects resemble those under egocentric questions, for which heading is irrelevant; the egocentric camera-reversal condition confirms that answers respond to camera changes. Camera replacement changes residuals only 1.250–1.830 times as much as heading replacement, far less than the 22–55-fold difference in answer effects. At InternVL3.5-8B layer 18, heading replacement even causes a larger residual perturbation (0.449 versus 0.385), yet its answer effect is about 20 times smaller. Residual perturbation magnitude alone does not explain the difference in answer sensitivity.

The fine-tuned heading classifier gives an effect of 1.000 under heading replacement, showing that this measurement detects a large effect when the model uses heading. Repeating the camera-versus-heading comparison on the log-odds scale gives effect ratios of 9.600–49.300. Together, the results show stronger answer dependence on camera-relative information than on reference heading under the tested allocentric questions.

F Testing Whether the Answer Mapping Rotates with Reference Heading

An answer tied to the reference object's coordinate system should shift with its heading as the target moves around it. We fit the model's yes/no response curves over target azimuth and test for the 180° phase shift required by heading reversal.

F.1 Models

We test all eight models in Appendix C.1. For each model, we separately ask about left, right, front, and behind in the reference object's coordinates.

F.2 Harmonic fitting and model selection

Each input contains one image and one binary allocentric question. We vary the target's azimuth relative to the camera while keeping the queried relation fixed, and record the answer margin, the yes logit minus the no logit. We repeat this measurement for left-facing and right-facing reference objects. After the target completes a full circle, the scene repeats, so the margin can be analyzed as a periodic function of azimuth. If $\hat{u}_\varphi$ is the queried direction and $\hat{r}(\theta)$ is the unit vector toward the target, their projection is $\hat{u}_\varphi^\top \hat{r}(\theta) = \cos(\theta - \varphi)$. Its sign determines the binary direction judgment in that coordinate system. We therefore fit the first harmonic to extract one sinusoidal cycle from the response curve.

At 36 distinct, uniformly spaced azimuths, we fit a linear combination of 1, $\cos\theta$, and $\sin\theta$ by least squares, then convert the two trigonometric coefficients into an amplitude and phase. The amplitude measures how strongly the answer margin varies with target position, and the phase locates that variation around the circle. Higher-frequency components that can be distinguished at these sampling points are orthogonal to the first harmonic and remain in the fitting residual. We interpret phase only when the amplitude is large enough and the output is not saturated near a constant answer. Because two curves separated by 180° can cancel when pooled, amplitude screening uses the mean of the amplitudes fitted to the two headings separately, $\bar{a} = (a_L + a_R)/2$. The amplitude a_0 obtained by fitting both headings together is reported separately in Table 23 and Figure 5(a).

We compare two fitting models. Model $\mathcal{M}_0$ uses one parameter set (a_0, φ, b) for both reference headings, where b is the constant offset. Model $\mathcal{M}_1$ fits an amplitude, phase, and offset to each heading separately. We measure goodness of fit as

$$\mathrm{RSS} = \sum_i (m_i - \hat{m}_i)^2, \quad \mathrm{TSS} = \sum_i (m_i - \bar{m})^2, \quad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}. \tag{17}$$

where m_i, $\hat{m}_i$, and $\bar{m}$ are the observed margins, fitted margins, and sample mean of the observed margins. A high R^2 indicates that the sinusoid explains the observed variation. The separate phase comparison measures the shift under heading reversal.

The Akaike information criterion is $\mathrm{AIC} = 2k - 2\ln\hat{L}$. Under Gaussian noise, its difference is (Akaike, 1974)

$$\mathrm{dAIC} = \mathrm{AIC}(\mathcal{M}_1) - \mathrm{AIC}(\mathcal{M}_0) = 2(k_1 - k_0) + n\ln\frac{\mathrm{RSS}_1}{\mathrm{RSS}_0}, \tag{18}$$

Here $k_0 = 3$ and $k_1 = 6$ are the parameter counts, and n is the number of observations. Because $\mathcal{M}_0$ is a special case of $\mathcal{M}_1$, fitting the two headings separately cannot increase the residual sum of squares. The Akaike information criterion (AIC) penalizes the additional parameters. Negative dAIC favors separate fits, but correct reference-frame binding additionally requires the phase shift in Eq. (7).

F.3 Data and resampling

For each model and queried relation, we collect unmodified answer margins at 36 equally spaced target azimuths for each reference heading. Each observation comes from one image and one binary allocentric question.

We repeatedly sample images with replacement, refit $\mathcal{M}_0$ and $\mathcal{M}_1$, and recompute dAIC. Across $B = 4000$ such bootstrap samples, we report the fraction with $\mathrm{dAIC} \le -10$, together with the value fitted to the original images. The resampling unit is the image. When this fraction is near 0.0500, random variation from the finite number of resamples has a Monte Carlo standard error of approximately 0.0035. We repeat the resampling for one model with four random seeds to assess its sensitivity to the sampled images.

F.4 Checking phase estimation on constructed response curves

We create two sets of periodic response curves whose phase relationships are specified in advance. In the first, reversing the reference heading leaves the response curve unchanged. In the second, it shifts the curve by half a circle, as required for correct binding. We add noise with standard deviation σ and apply the same fitting procedure used for model outputs (Table 21). Without noise, the recovered phase differences are 0° and 180°. For the correctly bound response, fitting the opposite curves together gives $a_0 = 0$, whereas fitting them separately gives $\bar{a} = 3$. Screening on pooled amplitude would therefore reject this valid response. This experiment verifies the phase estimator and amplitude screen on curves with known heading-dependent behavior.

F.5 Phase differences

The response curves are fitted separately for the two reference headings (Figure 3); the full comparison covers 32 combinations of eight models and four relations (Figure 5).

Table 21: Phase estimates for constructed curves with known behavior under heading reversal. The camera-frame response has identical curves for the two headings; the correctly bound response has curves separated by $180°$. Columns give recovered $|\Delta\varphi|$ after adding noise of standard deviation σ.

σ	Camera-frame response	Correctly bound response
0.0000	0.0000°	180.0000°
0.2500	0.3000°	179.7000°
0.5000	0.6100°	179.3000°
1.0000	1.2200°	178.7000°

For the right-relation question in Qwen2.5-VL-7B, the left-heading fit is $(a, \varphi, b) = (7.3384, 92.7170°, 1.3333)$ and the right-heading fit is $(7.3475, 93.0520°, 1.4913)$, with pooled $R^2 = 0.9554$. For the behind-relation question in InternVL3-8B, the corresponding fits are $(3.3676, 175.7910°, 1.6069)$ and $(3.4048, 175.4730°, 1.6163)$, with pooled $R^2 = 0.9046$. These examples show the similar phase under opposite reference headings; the tables below cover all combinations.

Table 22: Phase differences after excluding weak response curves at each listed amplitude threshold. Screening uses $\bar{a}$, the mean amplitude fitted separately to left-facing and right-facing reference objects. The count n is the number of retained combinations of model and queried relation; the final columns give their maximum and median absolute phase difference. At these four thresholds, the same combinations are retained when screening on the amplitude a_0 fitted to both headings together.

Amplitude threshold	n	Maximum $\|\Delta\varphi\|$	Median $\|\Delta\varphi\|$
$\bar{a} \geq 0.5000$	27	63.0700°	0.5400°
$\mathbf{\bar{a} \geq 1.0000}$	**19**	**1.9900°**	0.3500°
$\bar{a} \geq 2.0000$	15	1.9900°	0.3300°
$\bar{a} \geq 3.0000$	12	1.9900°	0.3400°

Table 23: The queried relation with the largest amplitude a_0 for each model. The amplitude and R^2 are obtained by fitting both reference headings together, while $|\Delta\varphi|$ compares their separate fits. Phase-based binding comparisons use responses with sufficient amplitude and unsaturated output; the other entries describe their fitted curves.

Model	Relation	a_0	$\|\Delta\varphi\|$	R^2
InternVL3-8B	Right	6.2210	0.5400°	0.9232
InternVL3.5-8B	Right	9.9960	0.6200°	0.8759
Qwen2.5-VL-7B	Right	7.3430	0.3300°	0.9554
LLaVA-OneVision-7B	Left	2.5530	0.3300°	0.8941
llava-1.5-7B	Right	1.3340	0.4000°	0.8755
idefics2-8B	Left	0.9070	0.4600°	0.5280
Phi-3.5-Vision	Right	3.0140	**1.9900°**	0.7538
Gemma-3-4B	Left	1.0060	1.2700°	0.2703

The 19 combinations with $\bar{a} \geq 1.0000$ cover seven models; no idefics2-8B combination qualifies (Table 22). Screening by pooled amplitude a_0 or by the mean separate-heading amplitude $\bar{a}$ happens to select the same combinations at all four thresholds in these model outputs. Each model's strongest-amplitude relation has a phase difference of at most $1.9900°$, with idefics2-8B excluded from the phase-based binding comparison for low amplitude and Gemma-3-4B excluded for output saturation (Table 23). Increasing the amplitude threshold from 0.5000 to 1.0000 reduces the maximum phase difference from $63.0700°$ to $1.9900°$; thresholds of 2.0000 and 3.0000 leave it unchanged. Larger phase differences occur among weaker response curves: $\mathrm{Spearman}(\log a_0, \log|\Delta\varphi|) = -0.5550$, and the slope after taking logarithms of both variables is -0.6680 (the simple phase-noise prediction is -1). All five differences above $8°$ occur at $a_0 < 1.0000$.

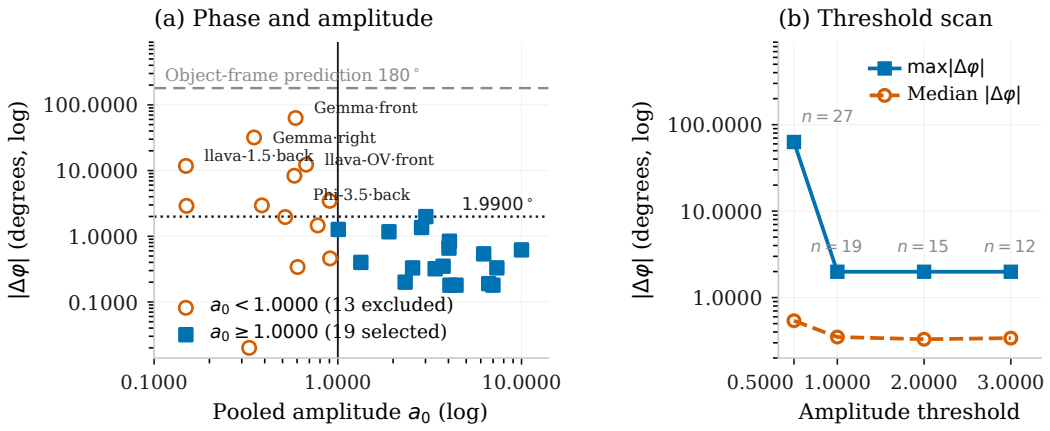

Figure 5: Phase difference and response amplitude for binary allocentric questions. (a) All 32 combinations of model and queried relation on logarithmic axes, using the amplitude a_0 fitted jointly to both reference headings. The horizontal dashed line marks the $180°$ shift required by heading reversal; the dotted line marks the maximum retained difference, $1.9900°$. The threshold $a_0 = 1.0000$ retains the same original-model combinations as $\bar{a} \geq 1.0000$. Labels identify the five differences above $8°$. The measured phase difference for Phi-3.5-Vision on the front relation is $0°$, displayed at $0.0200°$ because zero cannot be plotted on a logarithmic axis. (b) The four amplitude thresholds in Table 22, with retained counts n.

Weighting each phase difference by its amplitude still includes phases estimated from weak response curves. For example, the weighted difference for Gemma-3-4B reaches $21.7600°$, while its output remains saturated. The amplitude and saturation criteria also apply to the weighted phase statistic.

F.6 Goodness of fit and model selection

Table 24: Goodness of the shared sinusoidal fit and comparison with separate-heading fits, summarized over each model's four queried relations. Columns give median and minimum R^2, the number of relations with $R^2 < 0.8000$, minimum dAIC, and the largest bootstrap fraction with $\mathrm{dAIC} \leq -10$. The last two columns include all relations, including those excluded by the amplitude threshold.

Model	Median R^2	Min. R^2	Count $R^2 < 0.8000$	Min. dAIC	Max. tail frequency
Qwen2.5-VL-7B	0.9474	0.7451	1	+2.1930	0.0493
InternVL3-8B	0.9139	0.9001	0	+1.1660	0.0800
InternVL3.5-8B	0.8109	0.7245	2	+0.4030	0.0895
LLaVA-OneVision-7B	0.6387	0.3045	2	+0.3970	0.0920
llava-1.5-7B	0.5508	0.2559	2	−0.8540	0.1460
Phi-3.5-Vision	0.5365	0.1006	3	−4.2660	0.2915
idefics2-8B	0.5184	0.3445	4	+2.6010	0.0398
Gemma-3-4B	0.0334	0.0244	4	−4.6110	0.3125

For the 12 combinations of four relations and the three primary models, InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, every point estimate of dAIC is nonnegative, favoring the shared fit $\mathcal{M}_0$ (Table 24). The bootstrap fractions with $\mathrm{dAIC} \leq -10$ range from 0.0013 to 0.0895. Qwen2.5-VL-7B's fraction of 0.0493 is only 0.2000 Monte Carlo standard errors below 0.0500. For four InternVL3.5-8B resampling seeds, the fractions span 0.0895–0.1025. A cutoff of 0.1000 would classify this model differently across resampling seeds even though the phase differences remain small.

Fit quality also varies by queried relation. For LLaVA-OneVision-7B, llava-1.5-7B, and Phi-3.5-Vision, front/behind amplitudes are only 0.1500–0.6700, whereas left/right amplitudes are 1.0700–3.0100; the phases of the left/right responses are more interpretable. All four idefics2-8B amplitudes are weak (0.5200–0.9100). Gemma-3-4B's margin varies around $+9$ to $+11$, with a within-condition standard deviation of about 1.0000, but this variation follows target azimuth poorly. Its ratio of margin shifts for images whose labels remain unchanged to shifts for images whose labels change is 0.9500 (Appendix D); similar shifts occur whether or not heading reversal requires the answer to change. The other seven models answer yes on 0.4479–0.6253 of samples, whereas Gemma-3-4B does so on 0.9969. Its median absolute margin is 8.7500 and its margin range is $[-2.7500, 13.7500]$.

G Targeted Fine-Tuning and Generalization Tests

We report two fine-tuning experiments. Experiment 1 trains on some reference-object categories and tests on different categories using both reference headings. Experiment 2 trains with only one heading and evaluates object-category changes and heading changes separately. Additional comparisons vary the trainable layers and the loss term that preserves egocentric predictions.

G.1 Models

Experiment 1 uses InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B with three training settings each, giving nine runs. The layer and loss comparisons use the same models with four settings each, giving 12 runs. Experiment 2 evaluates three trained checkpoints: InternVL3.5-8B with random seeds 0 and 1, and llava-1.5-7B with seed 0. The llava-1.5-7B checkpoint supplies a different language-model family.

G.2 Four-choice evaluation metrics

For questions with four answer options, let $\mathcal{D} = \{\text{left}, \text{right}, \text{front}, \text{back}\}$ be the direction choices, with correct direction y_i and predicted direction $\hat{a}_i$ for example i among N evaluated examples. The mapping π returns the opposite direction, swapping left with right and front with back. We define

$$\text{acc} = \frac{1}{N}\sum_i \mathbf{1}[\hat{a}_i = y_i], \qquad p_{\text{opp}} = \frac{1}{N}\sum_i \mathbf{1}[\hat{a}_i = \pi(y_i)], \tag{19}$$

$$N_c = \sum_i \mathbf{1}[\hat{a}_i = c], \quad c \in \mathcal{D}, \qquad z(\hat{p}) = \frac{\hat{p} - 0.2500}{\sqrt{0.2500 \times 0.7500/N}}. \tag{20}$$

Accuracy is the fraction of correct directions, and the antipodal rate p_{opp} is the fraction of predictions exactly opposite the correct direction. The counts N_c show how often each direction is predicted and distinguish scene-dependent reversed answers from always choosing one option. The statistic z expresses a proportion's deviation from the four-choice chance level of 0.2500 in units of its binomial standard deviation.

G.3 Training objectives and answer probabilities

For the binary-question experiment, we combine a loss for answering allocentric questions, a loss for changing the answer appropriately after heading reversal, and a loss for preserving the pretrained model's egocentric predictions.

$$\mathcal{L} = \mathcal{L}_{\text{allo}} + \lambda_{\text{sens}}\,\mathcal{L}_{\text{sens}} + \lambda_{\text{ego}}\,\mathcal{L}_{\text{ego}}. \tag{21}$$

The three losses use probabilities computed differentiably from the final answer-position token scores. Let s_v denote the logit of token v, and let $\mathcal{T}_{\text{yes}}$ and $\mathcal{T}_{\text{no}}$ contain token variants corresponding to yes and no. We define

$$\delta = \ln \sum_{v \in \mathcal{T}_{\text{yes}}} e^{s_v} \;-\; \ln \sum_{v \in \mathcal{T}_{\text{no}}} e^{s_v}, \qquad p = \sigma(\delta), \tag{22}$$

Here σ is the sigmoid function. For each answer category, log-sum-exp combines the scores of its synonymous tokens. Their difference δ is the token-group version of the yes-minus-no answer margin m.

Because assigning the same training target to opposite-heading images can reward ignoring the queried spatial relation, the allocentric classification loss uses the geometric answer for each individual image. Let p_L, p_R be the predicted yes probabilities and $t_L, t_R \in \{0, 1\}$ the correct yes/no labels for the left-facing and right-facing images:

$$\mathcal{L}_{\text{allo}} = -\sum_{c\in\{L,R\}} \left[t_c \ln p_c + (1 - t_c) \ln(1 - p_c)\right]. \tag{23}$$

This loss sums over the pair. An uninformative predictor with $p_L = p_R = 0.5000$ has loss $2(-\ln 0.5000) = 1.3863$. Values below 1.3863 indicate fitting beyond this baseline.

The paired-sensitivity loss directly matches the probability difference to the label difference:

$$\mathcal{L}_{\text{sens}} = \left[(p_L - p_R) - (t_L - t_R)\right]^2. \tag{24}$$

When the two images have opposite correct labels, predicting the same probability for both gives a nonzero sensitivity loss, including when both probabilities equal 0.5000. This loss penalizes failure to change the answer in the direction required by heading reversal.

The egocentric preservation loss penalizes squared drift from the pretrained model:

$$\mathcal{L}_{\text{ego}} = \sum_{c\in\{L,R\}} (p_c^{\text{ego}} - \tilde{p}_c^{\text{ego}})^2, \tag{25}$$

where $\tilde{p}_c^{\text{ego}}$ is the pretrained model's probability, computed before fine-tuning and kept fixed as the target. This term penalizes changes to egocentric predictions while the model is trained on allocentric questions.

Experiment 2 uses two loss terms. For four-choice questions, the objective becomes

$$\mathcal{L}' = \mathcal{L}_{\text{dir}} + \lambda_{\text{ego}} \mathcal{L}'_{\text{ego}}, \qquad u_c = \ln \sum_{v\in\mathcal{T}_c} e^{s_v}, \qquad q = \text{softmax}(u), \tag{26}$$

Here $\mathcal{T}_c$ contains the first tokens of answer-option letters corresponding to direction $c \in \mathcal{D}$, and u_c combines their logits. The vector q contains the resulting option probabilities. The direction loss is $\mathcal{L}_{\text{dir}} = -\ln q_y$, where y is the correct direction. The preservation loss is $\mathcal{L}'_{\text{ego}} = |\mathcal{D}|^{-1} \sum_c (q_c^{\text{ego}} - \tilde{q}_c^{\text{ego}})^2$, where $\tilde{q}_c^{\text{ego}}$ is the pretrained option probability. We set $\lambda_{\text{ego}} = 0.5000$.

G.4 Training and testing on different reference-object categories

Experiment 1 uses COMFORT heading-reversal pairs. Each input contains one image and one binary spatial question. Allocentric questions supply the classification and pair-sensitivity losses, and egocentric questions supply the preservation loss. Reference-object categories are assigned wholly to either training or testing, so a test category does not appear in training. Evaluation includes both reference headings.

We compare three training settings that vary the relations used for training and the parameters allowed to change: low-rank adaptation (LoRA) for a single queried relation, LoRA trained jointly on all four relations, and full-parameter fine-tuning of the layers in the readout window. The readout window is the layer range identified by the activation-replacement analysis in Appendix H.

Training D is at least 0.8300 in every run, and the final allocentric classification loss $\mathcal{L}_{\text{allo}}$ ranges from 0.0000 to 0.5590, below the uninformative-prediction baseline of 1.3863 (Table 25). Nevertheless, every test accuracy is at most 0.4850. Most test D values are negative, meaning that the predicted probability changes in the opposite direction to the required label change after heading reversal.

Full-parameter fine-tuning gives final classification losses of 0.0000 for InternVL3-8B, 0.0000 for InternVL3.5-8B, and 0.0510 for Qwen2.5-VL-7B, but their test accuracies are only 0.0560, 0.0280, and 0.1110, respectively. In the comparison that increases the number of training reference categories from four to eight, the training fit improves while test D changes from $+0.0190$ to -0.2580. The observed failure to generalize persists despite successful fitting and the tested increases in trainable parameters and training categories.

Under this experimental setup, fine-tuning fits seen-reference examples but fails to learn an allocentric rule that transfers across reference categories and headings.

Table 25: Directed correctness D for nine training settings evaluated with binary allocentric questions. Reference-object categories in the test set do not occur in training. Single-relation LoRA and Four-relation LoRA train low-rank adapters on one relation or all four relations, respectively; Full fine-tuning updates all parameters in the selected layers. Each setting reports training D, test D, and test accuracy in parentheses.

Model	Single-relation LoRA		Four-relation LoRA		Full fine-tuning	
	Train D	Test D	Train D	Test D	Train D	Test D
InternVL3-8B	$+1.0000$	-0.2580 (0.3610)	$+0.9580$	-0.2880 (0.2930)	$+1.0000$	-0.1890 (**0.0560**)
InternVL3.5-8B	$+0.9970$	-0.2840 (0.3330)	$+0.9960$	$+0.0060$ (0.4850)	$+1.0000$	-0.6460 (**0.0280**)
Qwen2.5-VL-7B	$+1.0000$	$+0.0800$ (0.4440)	$+0.9530$	-0.1660 (0.1920)	$+0.8320$	-0.1120 (**0.1110**)

G.5 Layer-placement and preservation-loss ablations

We next vary the layers updated during fine-tuning and whether the loss preserves egocentric predictions. All questions in this comparison ask about the behind relation. For each model, we train four versions: LoRA in six layers before the readout window, in the window, or in six layers after it, each with $\lambda_{\text{ego}} = 1.0000$; and LoRA in the window with $\lambda_{\text{ego}} = 0$. The 12 runs use learning rate 0.0001, three epochs, rank 16, and $\lambda_{\text{sens}} = 1.0000$. The train/test assignments and random seeds are fixed across comparisons. The training reference objects are a car, a bench, and a bed; testing uses a chair and a dog.

After training, we apply the heading-reversal activation replacement from Appendix E under allocentric questions, separately for reference categories included in training and categories used only for testing. Its effect ϵ is the mean absolute yes-probability change defined in Eq. (5). We also use camera-reversal donors under egocentric questions to check whether the model retains sensitivity to camera-relative position. The final $\mathcal{L}_{\text{ego}}$ measures the change from pretrained egocentric predictions (Table 26).

Table 26: Twelve comparisons of trainable layer range and egocentric-preservation loss. Before and After denote the six layers preceding or following the readout window. Train ϵ and Test ϵ are the heading-reversal replacement effect ϵ under allocentric questions, evaluated on reference categories used in training and categories used only for testing, respectively. Ego. camera effect is the camera-reversal replacement effect under egocentric questions. Lower $\mathcal{L}_{\text{ego}}$ means less change from pretrained egocentric predictions.

Model	Layers	λ_{ego}	Train ϵ	Test ϵ	Ego. camera effect	$\mathcal{L}_{\text{ego}}$
InternVL3-8B	Before	1.0000	0.8260	0.0610	0.7250	0.0096
	Window	1.0000	0.7460	0.1240	0.6810	0.0105
	After	1.0000	0.8220	0.0420	0.6990	0.0197
	Window	**0.0000**	0.6260	0.2390	**0.0710**	**0.3840**
InternVL3.5-8B	Before	1.0000	0.9910	0.0010	0.8680	0.0320
	Window	1.0000	0.9950	0.0250	0.8490	0.0211
	After	1.0000	0.9950	0.3300	0.8160	0.0007
	Window	**0.0000**	0.9660	0.0880	**0.0430**	**0.6568**
Qwen2.5-VL-7B	Before	1.0000	0.9940	0.3660	0.8290	0.0070
	Window	1.0000	0.9940	0.0300	0.7940	0.0028
	After	1.0000	0.9960	0.0280	0.8140	0.0358
	Window	**0.0000**	0.9930	0.0040	**0.0950**	**0.9357**

With egocentric preservation enabled, the heading-replacement effect on training categories is approximately 0.7400–1.0000 across layer choices, including layers after the readout window (Table 26). Seven of the nine effects on test categories are only 0.0010–0.1240. Changing the trainable layer range does not resolve the observed failure to generalize.

Without the preservation loss, the heading-replacement effect on training categories still reaches 0.6260 for InternVL3-8B, 0.9660 for InternVL3.5-8B, and 0.9930 for Qwen2.5-VL-7B, but generalization remains unsuccessful. Across the three models, the camera effect under egocentric questions falls from the reported 0.6810–0.8680 range with preservation to 0.0430–0.0950 without it, while $\mathcal{L}_{\text{ego}}$ rises from 0.0007–0.0358 to 0.3840–0.9357. Removing the preservation term damages egocentric behavior without restoring generalization.

G.6 Separating generalization to new objects from heading reversal

Experiment 2 trains only on left-facing reference objects from seven categories: car, bed, bench, dog, duck, horse, and laptop. Chair and sofa appear only in testing. Fifteen images per reference category and relation give 420 training questions. Each question asks for an allocentric direction using four answer options mapped to front, back, left, and right. All three checkpoints use LoRA rank 16, scaling 32, learning rate 0.0001, 15 epochs, and $\lambda_{\text{ego}} = 0.5000$. LoRA updates layers 18–24 in InternVL3.5-8B and 11–16 in llava-1.5-7B. InternVL3.5-8B seed 0 is evaluated at 15 target azimuths; the other checkpoints are evaluated at 37.

We evaluate four groups while keeping the question format fixed (Table 27). Group 1 uses training reference categories at the training heading. Group 2 changes only the category, using chair and sofa at the training heading. Group 3 uses training categories with the heading reversed. Group 4 combines test categories with reversed heading. Comparing Groups 1 and 2 measures transfer to different objects at a fixed heading; comparing Groups 2 and 4 measures the effect of heading reversal while both groups use categories absent from training.

Table 27: Four-choice allocentric evaluation after training with left-facing reference objects only. Trained heading is the training orientation, and Reversed heading is its 180° reversal. Seen objects belong to training reference categories; Unseen objects belong to categories reserved for testing. Each checkpoint is tested in all four combinations, with N questions in each group. The rate p_{opp} counts answers exactly opposite the correct direction. It was measured only for the two zero-accuracy groups; n/m means not measured.

Checkpoint	Group	Heading	Object	N	Accuracy	p_{opp}
InternVL3.5-8B (seed 0)	1	Trained	Seen	420	1.0000	n/m
	2	Trained	Unseen	120	0.9750	n/m
	3	Reversed	Seen	420	**0.0000**	**0.9881**
	4	Reversed	Unseen	120	**0.0000**	**1.0000**
InternVL3.5-8B (seed 1)	1	Trained	Seen	1036	1.0000	n/m
	2	Trained	Unseen	296	0.9595	n/m
	3	Reversed	Seen	1036	**0.0000**	**1.0000**
	4	Reversed	Unseen	296	**0.0000**	**0.9730**
llava-1.5-7B (seed 0)	1	Trained	Seen	1036	1.0000	n/m
	2	Trained	Unseen	296	0.9865	n/m
	3	Reversed	Seen	1036	**0.0000**	**0.9961**
	4	Reversed	Unseen	296	**0.0000**	**0.9831**

All checkpoints achieve 1.0000 accuracy in Group 1 (Table 27). These are the training samples for InternVL3.5-8B seed 0, whereas $22/37$ of the evaluated azimuths were absent from training for the other checkpoints. In Group 2, accuracy remains 0.9595–0.9865, a maximum decrease of 0.0405. Reversing the heading reduces accuracy to 0.0000 both for training categories in Group 3 and for test categories in Group 4. Groups 2 and 4 isolate heading reversal most directly, because neither contains reference categories used in training.

Zero accuracy is below the four-choice chance level of 0.2500, with $z(\text{acc}) = -6.3200$ to -18.5800. The opposite-direction rates are 0.9730–1.0000, with $z(p_{\text{opp}}) = +18.9700$ to $+55.7500$ and permutation $p < 0.0002$ at 5,000 resamples. The errors almost always indicate the direction 180° from the correct answer. Because always predicting the same option could obscure the meaning of this pattern, we also inspect the recorded Group 3 prediction counts for InternVL3.5-8B seed 0. The four

direction counts are 108, 105, 103, and 104 ($\chi^2 = 0.1300$, three degrees of freedom), despite zero correct answers. Its predictions vary across scenes but are reversed relative to the required answer.

These results support a learned rule that maps target azimuth and the queried relation to a direction but does not update that mapping when the reference heading changes. Because training uses a single heading, reversing that heading changes the correct geometric answer by 180° while leaving the learned mapping largely unchanged.

The four-group design separates object generalization from heading generalization, which Experiment 1 tests together. Both experiments test reference categories absent from training, but Experiment 1 evaluates both headings and requires correct binding across headings. Group 2 of Experiment 2 measures object generalization at the fixed training heading. Changing object category reduces accuracy by at most 0.0405, whereas reversing heading reduces accuracy from 1.0000 to 0.0000.

H Localizing the Readout Window

Image-token replacement and answer-position attention probe different stages of information use. We locate the layers beyond which image-token replacement has little effect, then ask whether attention at the answer position uses that information within the same range or later.

H.1 Localization procedure

Using one image and one binary allocentric question per input, we repeat the camera-reversal replacement from Appendix E at successive layers. At layer ℓ, we replace all recipient residual vectors at image-token positions with the vectors obtained from the paired image viewed from the opposite camera azimuth, then continue the forward pass. The resulting mean absolute yes-probability change is $\epsilon(\ell)$ in Eq. (5).

We identify the layer range from the shape of this effect curve. Early layers show a roughly constant effect near 0.600–0.750, followed by a decline toward zero. Replacement before this decline strongly changes the answer, whereas replacement after it has little effect. We call the declining range the **readout window**. The window locates the decline in answer sensitivity to changes at image-token positions.

H.2 Models and data

We measure detailed effect curves for InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B, and apply the same successive-layer replacement procedure to identify the window ranges for all eight models.

The camera-reversal pairs are the same as in Appendix E: the camera moves by 180° while the scene, reference heading, and world positions remain fixed. We repeat the measurement with heading-reversal pairs, where the camera and target position remain fixed, as a comparison. Both use the same binary allocentric question for the donor and recipient images and measure the recipient's yes probability.

H.3 Effects across depth

We sample seven representative locations along each model: visual-encoder output, language-model input, the early region of roughly constant effect, the start and middle of the declining range, a layer with a near-zero effect, and the end of the declining range. The window boundaries specify the layer range used in subsequent interventions (Table 29).

Camera-replacement effects remain nearly constant through the early layers, then decline sharply within the readout window in all three models (Table 28; Figure 6). For example, the effect in InternVL3-8B falls from 0.731 at the window start to 0.040 at layer 17. By the window end, camera effects are only 0.012–0.067 across the three models. Heading effects remain small throughout (0.010–0.050), overlapping the 0.018–0.041 range under egocentric questions in Appendix E. The pronounced decline with depth is specific to camera replacement in this comparison.

Table 28: Mean absolute changes in yes probability after replacing image-token representations with those from camera-reversal donors under allocentric questions. LM input is the language-model input. The plateau is the early range where effects remain approximately constant; the window is the later range where they decline. The final column gives effects from heading-reversal replacements evaluated in the same experiment.

Site	InternVL3-8B	Qwen2.5-VL-7B	InternVL3.5-8B	Heading control
Visual output	0.743	0.673	0.591	0.010–0.050
LM input	0.743	0.673	0.591	0.010–0.050
Pre-window plateau	~ 0.740	~ 0.670	~ 0.580	0.010–0.050
Window start	0.731 ($\ell = 14$)	0.660 ($\ell = 14$)	0.508 ($\ell = 18$)	Near noise level
Window middle	0.481 ($\ell = 16$)	0.363 ($\ell = 16$)	0.172 ($\ell = 20$)	Near noise level
Near-zero effect	0.040 ($\ell = 17$)	0.094 ($\ell = 17$)	0.056 ($\ell = 21$)	Near noise level
Window end	0.029 ($\ell = 19$)	0.067 ($\ell = 19$)	0.012 ($\ell = 24$)	Near noise level

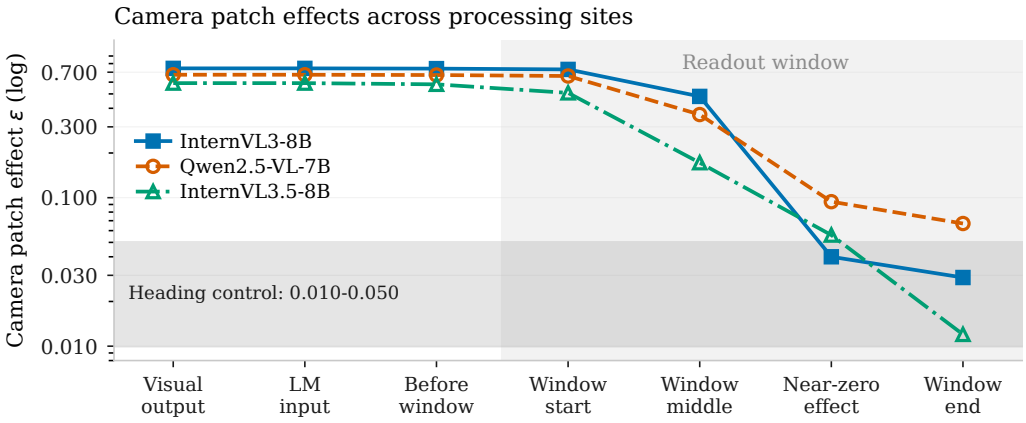

Figure 6: Mean absolute yes-probability changes after camera-reversal replacement at the seven locations in Table 28, plotted on a logarithmic effect axis. The four locations in the declining range are layers 14, 16, 17, and 19 for InternVL3-8B and Qwen2.5-VL-7B, and layers 18, 20, 21, and 24 for InternVL3.5-8B. The gray band gives effects from heading-reversal replacements evaluated in the same experiment.

H.4 Window versus later answer readout

We next modify attention outputs at the answer position instead of image-token representations. Starting at the beginning of the readout window, we replace the attention output at each tested layer with its mean over the dataset. We progressively extend the modified range to deeper layers, ending at layer ℓ, and measure the reduction in camera-induced answer change:

$$\mathrm{drop}(\ell) = 1 - \frac{\epsilon^m(\ell)}{\epsilon_0^m}, \tag{27}$$

where $\epsilon^m(\ell)$ is the camera-induced answer effect after replacing attention outputs from the window start through layer ℓ, and ϵ_0^m is the corresponding effect without attention replacement. The superscript m denotes the answer-margin scale, the yes logit minus the no logit. A larger $\mathrm{drop}(\ell)$ means that replacing attention outputs through layer ℓ removes more of the answer's camera sensitivity.

Because setting attention outputs to zero can change their overall scale, we instead replace them with dataset means, removing example-specific content while reducing this scale change.

Table 29: Layer ranges where the effect of replacing image-token representations declines. LM layers gives the total number of language-model layers, and Window gives the corresponding range of language-model layer indices, not visual-encoder layer indices. The final column identifies the visual encoder used by each model; the Gemma-3-4B range is approximate.

Model	LM layers	Window	Visual encoder
InternVL3-8B	28	$\ell = 14$–19	InternViT
InternVL3.5-8B	36	$\ell = 18$–24	InternViT
Qwen2.5-VL-7B	28	$\ell = 14$–19	Qwen-ViT
LLaVA-OneVision-7B	28	$\ell = 14$–19	SigLIP
llava-1.5-7B	32	$\ell = 11$–16	CLIP
Phi-3.5-Vision	32	$\ell = 16$–22	CLIP family
idefics2-8B	32	$\ell = 11$–16	SigLIP + Perceiver
Gemma-3-4B	34	$\ell = 12$–18 (approx.)	SigLIP

Table 30: Replacing answer-position attention outputs with their dataset means across successive layer ranges. Baseline camera and heading values are the magnitudes of yes-minus-no margin changes caused by each image transformation without attention replacement. Through window and Through final layer give $\mathrm{drop}(\ell)$ when replacement ends at the readout-window end or the final layer. Timing indicates whether most camera sensitivity is removed within the window, later, or across both ranges.

Model	Baseline camera	Baseline heading	Through window	Through final layer	Timing
InternVL3-8B	1.129	0.044	$+0.100$	$+0.980$	Later
InternVL3.5-8B	3.240	0.167	$+0.090$	$+0.990$	Later
Qwen2.5-VL-7B	1.898	0.090	$+0.110$	$+0.990$	Later
Phi-3.5-Vision	0.804	0.115	$+0.670$	$+0.960$	Mixed
Gemma-3-4B	0.354	0.221	$+0.210$	$+0.810$	Mixed
llava-1.5-7B	0.408	0.058	$+0.840$	$+0.910$	Within window
idefics2-8B	0.292	0.087	$+0.670$	$+0.570$	Within window

Extending attention replacement through the deepest layer removes 0.570–0.990 of camera sensitivity across seven models (Table 30). Before replacement, camera-induced margin changes span 0.292–3.240, while heading-induced changes stay below 0.300. In the three primary models, replacement only through the image-replacement window removes at most 0.110 of camera sensitivity, compared with 0.980–0.990 when extended through the deepest layer: their main answer-position use of camera information occurs later. For llava-1.5-7B and idefics2-8B, most of the reduction already occurs by the window endpoint. Phi-3.5-Vision and Gemma-3-4B show substantial additional reductions after that point. The timing classification depends on this later change: idefics2-8B and Phi-3.5-Vision both give 0.670 at the window endpoint, but extending replacement changes the reduction to 0.570 and 0.960, respectively.

I Tracing Camera Information Through the Model

We follow the camera dependence measured in Appendix E from visual-encoder outputs, through updates to the residual stream, to attention at the answer position. Inputs use paired images and the same binary allocentric question. Across experiments, we vary the replaced positions, layers, or components, then measure the resulting answer change. Probability effects use Eq. (5); the normalized comparison across layers additionally uses the log-odds measurement described below.

I.1 Models

InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B are the primary models for detailed layer and component measurements. Supplementary tests of residual updates extend to seven models, and

visual-encoder comparisons cover InternViT, Qwen-ViT, and SigLIP. Subspace analyses use the three primary models, with the comparison of a fixed basis across layers reported for InternVL3-8B and Qwen2.5-VL-7B.

I.2 Data and splits

We reuse the heading-reversal and camera-reversal pairs in Appendix E. The subspace analysis additionally uses 1,440 camera-reversal pairs, where the camera changes by 180° while the scene, object headings, and world positions remain fixed. Each intervention copies activations from one image in a pair into the forward pass for the other image, with the allocentric question unchanged.

Because fitting a subspace on its evaluation examples could reproduce their effects through overfitting, we learn subspaces on training pairs and evaluate them on separate test pairs. In five-fold cross-validation, four subsets of pairs are used for fitting and the remaining subset for testing, with each subset used as the test set in turn.

I.3 Camera information in visual-encoder outputs

We replace visual-encoder outputs for the recipient image with those of its camera-reversal donor before those outputs enter the language model. We compare this with replacement at the language-model input, using the same paired images and binary allocentric questions. InternVL models permit replacement before visual projection; Qwen2.5-VL permits replacement before its visual-token merger. The effects are 0.7430 for InternVL3-8B, 0.6730 for Qwen2.5-VL-7B, and 0.5910 for InternVL3.5-8B, equal to their language-model-input effects (Table 31). Repeating the procedure with heading-reversal donors gives effects of only 0.0100–0.0500. Camera-dependent answer effects are already present in the visual-encoder outputs.

Table 31: Mean absolute yes-probability changes under allocentric questions after replacing image representations at two locations. The donor image differs in camera azimuth. For InternVL models, visual output denotes the raw visual blocks before projection into language-model coordinates; for Qwen2.5-VL, it denotes the accessible output before visual-token merging. LM input denotes the language-model input. The final row uses heading-reversal donors in the same experiment.

Patch site	InternVL3-8B	Qwen2.5-VL-7B	InternVL3.5-8B
Visual output, before projection/merging	0.7430	0.6730	0.5910
LM input	0.7430	0.6730	0.5910
Heading control	0.0100	0.0200	0.0500

I.4 Contributions of individual and combined residual updates

We next distinguish replacing the accumulated residual representation from replacing only the update added by a particular component. For the same image pair and allocentric question, we copy the donor output of one attention head, one complete attention module, or one feed-forward module into the recipient at the corresponding layer and image-token positions. Earlier accumulated residual values remain unchanged. We then replace outputs jointly across a range of layers, testing attention updates, feed-forward updates, or both together. Heading-reversal donors provide a comparison with the camera-reversal donors. To compare a joint replacement with replacing the complete image-token residual, we define

$$\mathrm{rec} = \frac{\epsilon_{\mathrm{joint}}}{\epsilon_{\mathrm{seg}}}, \tag{28}$$

where $\epsilon_{\mathrm{joint}}$ is the mean absolute yes-probability change after jointly replacing component updates, and ϵ_{seg} is the effect of replacing the complete image-token residual representation. rec measures how much of that reference effect is reproduced by the joint component intervention.

Table 32: Mean absolute changes in yes probability after copying camera-reversal donor outputs under allocentric questions. Best in window is the maximum over individual heads, complete attention modules, and feed-forward modules within layers 14–19 for InternVL3 and Qwen, or 18–24 for InternVL3.5. Best in window and Window joint each use 24 pairs; the latter replaces attention and feed-forward updates together. Full-segment joint uses 40 pairs and replaces both update types from layer 1 through the window end. Recovery is rec, the full-segment effect divided by the complete image-token residual replacement effect.

Model	Best in window	Window joint	Full-segment joint	Recovery
InternVL3-8B	0.0347	0.4290	0.7200	0.9390
Qwen2.5-VL-7B	0.0437	0.2730	0.4240	0.6020
InternVL3.5-8B	0.0630	0.2430	0.5710	0.8250

Joint replacement within the readout window changes answers more than any individual component in all three models (Table 32). For InternVL3-8B, the joint effect is 0.4290, compared with 0.0347 for the strongest individual component. Joint replacement from layer 1 through the window end gives effects of 0.4240–0.7200 in the separate 40-pair evaluation. The larger joint effects are consistent with contributions distributed across components, with other components partly compensating when only one is changed.

Joint language-model updates reproduce 0.6020–0.9390 of the complete residual-replacement effect. The unrecovered fraction is consistent with a contribution from information already present at the language-model input. These ratios quantify recovery under joint intervention, including interactions between carried representations and later updates.

I.5 Comparing a fixed subspace with a subspace fitted at each layer

For InternVL3-8B and Qwen2.5-VL-7B, we examine more precisely where the effect of a particular set of residual directions declines. This comparison uses 240 image pairs from the basketball-and-bicycle configuration and three random seeds. We learn a 32-dimensional subspace from training pairs at layer 14, then keep its basis fixed while replacing only donor-recipient differences projected into this subspace at layers 12–18. We compare this with a subspace fitted separately at each tested layer. The first analysis asks whether the same directions continue to reproduce the effect; the second asks whether the information can still influence answers through different directions at that layer. Complete residual replacement, no replacement, and replacement within a random subspace of the same dimension provide comparisons. We compute

$$\mathrm{norm}(\ell) = \frac{\epsilon^{\log}_{\mathrm{sub}}(\ell) - \epsilon^{\log}_{0}(\ell)}{\epsilon^{\log}_{\mathrm{full}}(\ell) - \epsilon^{\log}_{0}(\ell)}, \qquad \mathrm{decay}(\ell \to \ell{+}1) = \frac{\Delta\mathrm{norm}/\mathrm{norm}}{\Delta\mathrm{gap}/\mathrm{gap}}, \tag{29}$$

Here $\epsilon^{\log}_{\mathrm{sub}}(\ell)$, $\epsilon^{\log}_{\mathrm{full}}(\ell)$, and $\epsilon^{\log}_{0}(\ell)$ are the answer effects when replacing components in the subspace, replacing the complete image-token residual, and making no replacement at layer ℓ, respectively. The superscript $\log$ denotes the answer log-odds scale. The difference $\mathrm{gap} = \epsilon^{\log}_{\mathrm{full}} - \epsilon^{\log}_{0}$ is the difference between complete replacement and no replacement at that layer. The normalized quantity $\mathrm{norm}(\ell)$ measures the fraction of this difference reproduced by the fixed subspace. The quantity decay compares the fractional decrease in this subspace effect between adjacent layers with the fractional decrease in the complete-replacement effect. It distinguishes loss of influence through the selected directions from the general reduction in the effects of interventions at deeper layers. Because a small denominator makes the fraction unstable, we restrict interpretation to layers with $\mathrm{gap} \geq 0.2000$.

Relative decay peaks between layers 15 and 16, at 2.7100 for InternVL3-8B and 4.2000 for Qwen2.5-VL-7B, compared with 1.2800 and 0.8500 between layers 14 and 15 (Table 33). The separately obtained causal peaks at layers 16 and 15 fall within the same range. Reducing the subspace dimension from 32 to 16 preserves the peak location and changes its value from 4.2000 to 4.0400, although the normalized fractions decrease by 0.1600–0.2400. At layer 16, the fixed basis reproduces

about 0.5200 of the complete-replacement effect in both models, whereas a basis fitted at that layer reproduces 0.8300 for InternVL3-8B and 0.7900 for Qwen2.5-VL-7B. The intervention effect can be recovered more strongly through different directions at the later layer, indicating a change in the subspace through which the information affects answers.

Table 33: Replacing residual components along the same layer-14 basis at successive layers. Range is the complete-replacement effect minus the no-replacement effect, $\epsilon^{\log}_{\text{full}} - \epsilon^{\log}_{0}$, measured using answer log-odds. Fixed fraction is $\text{norm}(\ell)$ in Eq. (29); Local fraction uses a basis fitted separately at the tested layer. Local − Fixed is Local fraction minus Fixed fraction. Layers 17–18 have Range below 0.2000, so the fractions are not interpretable, denoted n/a.

ℓ	InternVL3-8B				Qwen2.5-VL-7B	
	Range	Fixed fraction	Local fraction	Local − Fixed	Range	Fixed fraction
12	0.8500	0.9804	0.9804	0.0000	0.8080	0.9175
13	0.8460	0.9704	0.9655	−0.0049	0.7960	0.9476
14	0.8460	0.9606	0.9606	0.0000	0.7960	0.9529
15	0.6960	0.7425	0.8144	+0.0719	0.5710	0.7226
16	0.6170	0.5135	0.8311	+0.3176	0.5330	0.5234
17	0.0170	n/a	n/a	n/a	0.0170	n/a
18	0.0040	n/a	n/a	n/a	−0.0040	n/a

I.6 Contributions of attention heads at the answer position

Appendix H.4 shows that, for the three primary models, attention replacement must extend through the deepest layers to remove 0.9800–0.9900 of camera sensitivity. We next measure each head's contribution to the answer change under camera reversal, rank the heads by this contribution, and replace the outputs of increasing numbers of top-ranked heads at the answer position with their dataset means. Random sets containing the same number of heads provide a comparison. Inputs remain the paired images with binary allocentric questions.

Eight selected heads remove 0.3800–0.5900 of camera sensitivity, far more than eight randomly selected heads (0.0100–0.0200; Table 35). Increasing the selected set to 32 heads removes more sensitivity, but removing nearly all measured heads is required to reach 0.9700–0.9900.

We also evaluate whether head contributions are larger for camera reversal than heading reversal, whether replacing head outputs changes the answer, and whether the heading direction aligns with the final yes/no readout axis. In the latter half of the model, camera-to-heading contribution ratios range from 3.7000 to 18.9000 across six models and exceed random controls (Table 34). Absolute heading-to-readout cosines are only 0.0001–0.0132 where measured. Six models satisfy all three properties in the broader tests; the geometric property is supported in seven models.

The selected heads remove more camera sensitivity than equally sized random sets, so contributions are not uniform across heads. Nevertheless, removing nearly all sensitivity requires many heads. Camera sensitivity is distributed unevenly across many contributing heads.

I.7 Residual directions that reproduce the camera-replacement effect

For each image-token position, we learn directions of variation in the donor-minus-recipient residual vectors on training camera-reversal pairs. On separate test pairs, we project the donor-recipient difference onto the first k directions and add only that projected difference to the recipient residual before continuing the forward pass. Varying k measures how many residual-channel directions are needed to reproduce the complete-replacement effect. We compare these interventions with random isotropic Gaussian perturbations scaled to the same norm and with perturbations whose residual-channel coordinates are randomly reordered while their norm is preserved (Table 36; Figure 2). All conditions use binary allocentric questions.

Replacing components along only 16–32 selected directions reproduces 0.8000 of the complete camera-replacement effect, despite residual vectors having thousands of dimensions (Table 36).

Table 34: Numerical results for the attention-head and representation comparisons. Camera/heading ratio divides the camera-induced contribution by the heading-induced contribution in the latter half of the model. The cosine compares the logistic-regression heading direction $\hat{w}_{\text{probe}}$ with the yes/no readout axis g, using 296 final-answer-position samples per model. The model coverage is specified for each measurement; n/m denotes an unmeasured entry. The separate behavioral test records whether replacing head outputs changes answers.

Model	Camera/heading ratio	$\lvert\cos(\hat{w}_{\text{probe}}, g)\rvert$
InternVL3-8B	18.9000	0.0061
InternVL3.5-8B	9.2000	0.0001
Qwen2.5-VL-7B	15.2000	0.0067
LLaVA-OneVision-7B	n/m	0.0132
Gemma-3-4B	n/m	0.0097
llava-1.5-7B	4.2000	n/m
idefics2-8B	4.4000	n/m
Phi-3.5-Vision	3.7000	n/m

Table 35: Replacing selected attention-head outputs at the answer position with their dataset means. Columns give the fraction of camera sensitivity removed after replacing the 8 highest-contributing heads, the 32 highest-contributing heads, or all heads in the listed layer range. For randomly selected sets of the same size, reductions are 0.0100–0.0200 at $n_{\text{heads}} = 8$ and 0.1400–0.2400 at $n_{\text{heads}} = 32$; random controls are summarized by their ranges across models.

Model	Layer band	Camera-sensitivity reduction		
		$n_{\text{heads}} = 8$	$n_{\text{heads}} = 32$	All
InternVL3-8B	$\ell = 20$–28	0.3800	0.4800	0.9700
Qwen2.5-VL-7B	$\ell = 20$–28	0.5900	0.7100	0.9800
InternVL3.5-8B	$\ell = 25$–36	0.5300	0.8300	0.9900

Complete replacement is 5–8 times stronger than each model's stronger random control at the same perturbation norm; heading-reversal effects remain only 0.0100–0.0240. The selected camera subspace is nearly orthogonal to the final yes/no readout axis g but still changes the answer. Its substantial intervention effect despite weak direct alignment indicates that downstream computation transforms the intervened information.

J Reference-Frame Instructions, Object Localization, and Depth

Allocentric judgments require understanding the requested reference frame and locating the objects before applying the reference object's coordinates. The following experiments examine instruction recognition across wordings, compare egocentric and allocentric localization, and test depth decoding with different separations between training and test samples.

J.1 Recognizing reference-frame instructions across different wordings

A probe that distinguishes egocentric from allocentric instructions may rely on particular words rather than their meaning. We test seven models using 40 COMFORT images containing a car and a reference object, with three sets of differently worded instructions. The expressions specifying the reference frame do not overlap across sets. Because merely mentioning an external object could otherwise identify the allocentric condition, both kinds of instruction name the same external object and ask about the same spatial relation.

Using representations from instruction set A, we calculate $\hat{v}_A = \text{mean}(\text{allo}_A) - \text{mean}(\text{ego}_A)$, the difference between the mean residual representations for allocentric and egocentric instructions. We

Table 36: Effects of replacing residual components under allocentric questions. Full patch copies all image-token residual components from the camera-reversal donor. Gaussian and Shuffled direction apply random Gaussian perturbations or randomly reorder residual-channel coordinates, respectively, at the same perturbation norm. Heading control uses a heading-reversal donor. All effect columns report mean absolute yes-probability change. The dimension $d_{0.8000}$ is the number of selected channel directions needed to reproduce 0.8000 of the complete camera-replacement effect on test pairs.

Model	Full patch	Gaussian	Shuffled direction	Heading control	$d_{0.8000}$
InternVL3-8B	0.7310	0.0930	0.0600	0.0100	16
Qwen2.5-VL-7B	0.6600	0.1110	0.0720	0.0180	32
InternVL3.5-8B	0.5080	0.0840	0.1070	0.0240	32

keep this direction fixed and use it to classify the requested reference frame in a different wording set $\mathcal{B}$. The classification accuracy is

$$\tau = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathbf{1}\left[\operatorname{sign}\left(\hat{v}_A^\top h_i - t\right) = c_i \right], \tag{30}$$

where h_i is the residual representation of sample i in the new wording set, $c_i \in \{-1, +1\}$ identifies its egocentric or allocentric instruction, t is the classification threshold, and $\mathbf{1}[\cdot]$ equals one for a correct classification and zero otherwise. The direction remains fixed, but the threshold is recalibrated on the target wording set. Thus τ measures transfer of the separating direction after wording changes, with threshold recalibration on the target wording set.

We categorize $\tau \geq 0.6000$ as transfer, intermediate values between 0.2000 and 0.6000 as partial transfer, and $\tau \leq 0.2000$ as dependence on the original wording. We compare the results with classifiers fitted after shuffling reference-frame labels and with a control that tests wording alone. A further control names an external object without requesting its viewpoint. Its score is the fraction of these instructions classified as allocentric; a high score indicates that mentioning another object is being mistaken for requesting that object's reference frame.

Table 37: Reference-frame classification after changing instruction wording. τ is the accuracy in Eq. (30), with a fixed class-mean direction and a threshold recalibrated for the target wording set. External reference is the fraction of instructions classified as allocentric when they mention another object but do not request its viewpoint. The wording-only control tests classification attributable to wording. Interpretation combines τ with the shuffled-label and external-object-mention controls.

Model	τ	External reference	Interpretation
llava-1.5-7B	0.9730	0.0000	Transfers
Phi-3.5-Vision	0.9620	0.3333	Transfers
InternVL3-8B	0.8240	0.0000	Transfers
Qwen2.5-VL-7B	0.8230	0.0000	Transfers
InternVL3.5-8B	0.5790	0.0000	Partial
idefics2-8B	0.9530	**0.9833**	**Mention confound**
Gemma-3-4B	0.6790	0.6389	**Inconclusive**
Wording-only control	*0.5300*	n/a	n/a

Without threshold recalibration, InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B score 0.2180, 0.1780, and 0.4720, respectively, although each reaches 1.0000 when training and evaluation use the same wording set (Table 37). After recalibration, four models score 0.8230–0.9730 and InternVL3.5-8B scores 0.5790; these five exceed their shuffled-label baselines by 0.4500–0.7500. Their external-reference scores are zero except for Phi-3.5-Vision (0.3333). These comparisons support transfer across wording in four models and partial transfer in InternVL3.5-8B. In idefics2-8B, $\tau = 0.9530$ accompanies an external-reference score of 0.9833, so object mention provides an alternative explanation. Gemma-3-4B is inconclusive because no layer jointly reaches $\tau \geq 0.6000$ and exceeds

its shuffled-label baseline by at least 0.2500. Layers 9 and 34 have $\tau = 0.9000$ and 0.8500 but shuffled-label accuracies of 0.9100 and 0.7400; layer 13 exceeds the baseline by the required amount but has $\tau = 0.2570$.

J.2 Recognizing and locating two objects

We give all eight models SpinBench images and questions about the spatial relation between two named objects. Egocentric questions require the model to identify both objects and judge their positions relative to the observer. Allocentric questions additionally require a judgment in the reference object's coordinate system. We report egocentric accuracy on depth and left/right relations, overall accuracy under each reference frame, and a comparison restricted to left/right questions under the two reference frames.

Table 38: Accuracy on SpinBench questions involving two named objects. The first two columns evaluate egocentric depth and left/right judgments. Overall ego. and allo. denote accuracy across all egocentric and allocentric questions, respectively. The final two columns restrict the comparison to left/right questions: Grounding is egocentric accuracy, and Transformation is allocentric accuracy. These columns distinguish locating the objects relative to the observer from answering in the reference object's coordinate system.

Model	Grounding		Overall ego.	Overall allo.	Matched left/right	
	Depth	Left/right			Grounding	Transformation
InternVL3-8B	0.9605	0.9755	0.9387	0.5379	0.9755	0.5566
InternVL3.5-8B	0.9671	0.9301	0.9418	0.5724	0.9301	0.5294
Qwen2.5-VL-7B	0.9539	0.9231	0.9340	0.5276	0.9231	0.5430
LLaVA-OneVision-7B	0.9474	0.9021	0.9230	0.4862	0.9021	0.5158
Phi-3.5-Vision	0.7961	0.9266	0.8097	0.4552	0.9266	0.5339
llava-1.5-7B	0.6974	0.7448	0.6887	0.4379	0.7448	0.4842
idefics2-8B	0.7829	0.6993	0.6950	0.5276	0.6993	0.5430
Gemma-3-4B	0.6513	0.7762	0.6871	0.4759	0.7762	0.5294

All eight models answer SpinBench left/right questions more accurately in the egocentric frame (Table 38). The gap ranges from 0.1563 for idefics2-8B to 0.4189 for InternVL3-8B and also appears across the full set of relations. Requiring the reference object's coordinate system adds difficulty beyond the localization shared by the two question types.

J.3 Depth perception

We use 152 COMFORT questions about the depth relation between two objects for each of InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B. Probes predict the allocentric depth label from residual representations at the language-model input and at middle layer 17. At each location, we compare a difference-of-means (DoM) projection, logistic regression, and a nonlinear probe. All samples with the same dataset-provided group identifier remain together in either training or testing; the groups are never divided between the two. Each probe is also trained and evaluated after shuffling the labels, with the same data separation. Two additional measurements provide comparisons: a classifier predicts the same depth labels from the objects' two-dimensional image centers, and the model directly answers egocentric depth questions. The first checks whether the labels can be recovered from image coordinates; the second checks whether the model can answer the corresponding observer-centered questions.

Allocentric depth probes fail to exceed their corresponding shuffled-label controls in 17 of the 18 combinations of model, representation location, and probe (Table 39). The exception is InternVL3-8B's nonlinear probe at the middle layer, which scores 0.6161 versus 0.5753. Object-center coordinates provide a much stronger predictor: logistic regression reaches 0.9905 versus 0.5977 after shuffling. The three models also answer related egocentric depth questions with accuracy 0.8487–0.9803. Usable positional cues are present, but the tested residual probes usually fail to recover allocentric labels beyond their controls.

Table 39: Depth-label decoding on 152 questions per model. Input and Middle indicate residual representations at the language-model input and at layer 17. DoM is a projection along the difference between class means; Logistic is logistic regression. Observed uses the original labels, and Shuffled uses shuffled labels with the same samples, train/test groups, and probe selection procedure. The position baseline replaces residual features with two-dimensional object centers. Direct egocentric depth accuracy is 0.9803 for InternVL3-8B, 0.8487 for InternVL3.5-8B, and 0.9342 for Qwen2.5-VL-7B.

Model	Site	DoM		Logistic		Nonlinear	
		Observed	Shuffled	Observed	Shuffled	Observed	Shuffled
InternVL3-8B	Input	0.4300	0.5553	0.4642	0.5855	0.4619	0.5948
InternVL3-8B	Middle	0.5158	0.5622	0.5434	0.5760	0.6161	0.5753
InternVL3.5-8B	Input	0.3949	0.5239	0.5190	0.5960	0.4478	0.5205
InternVL3.5-8B	Middle	0.4857	0.5699	0.5000	0.6822	0.5160	0.6124
Qwen2.5-VL-7B	Input	0.4395	0.5772	0.5020	0.6062	0.4746	0.6005
Qwen2.5-VL-7B	Middle	0.4863	0.5805	0.5180	0.6222	0.5852	0.6165
Position baseline	n/a	0.9745	0.4513	0.9905	0.5977	0.9745	0.5022

J.4 Depth decoding under different training and test separations

Using a larger set of 400 questions per model, we compare three ways to evaluate the depth probes. First, individual question samples are randomly assigned to training and testing without keeping related questions together. Second, the correspondence between image features and labels is shuffled before applying that same random division. Third, all questions asking about the same spatial relation are assigned to the same group, and the test relation does not appear in training. We evaluate egocentric and allocentric labels using the same samples, probes, and separation rules. The last condition tests whether the learned depth information transfers to a spatial relation absent from probe training.

With individual questions randomly divided, allocentric depth decoding is high (linear accuracy 0.9325–0.9575; Table 40), but testing on a relation absent from training reduces it to near chance, as does shuffling labels. Egocentric decoding survives the same exclusion of the test relation almost intact (≥ 0.9916), revealing an asymmetry in transfer across queried relations.

Table 40: Depth-probe accuracy on 400 questions per model under three training/test conditions. Ungrouped randomly divides individual questions; Image–label shuffled first breaks the correspondence between features and labels; Relation grouped tests on a relation absent from training. Allo. and Ego. denote allocentric and egocentric labels. Linear is the better score from DoM projection and logistic regression; Nonlinear uses the nonlinear probe. Values are the highest accuracies across the tested layers. Both reference frames use the same samples, probes, and separation rules.

Frame	Condition	InternVL3-8B		InternVL3.5-8B		Qwen2.5-VL-7B	
		Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
Allo.	Ungrouped	0.9500	0.9925	0.9325	0.9750	0.9575	0.9975
	Image–label shuffled	0.5275	0.5100	0.5175	0.5075	0.5100	0.4900
	Relation grouped	0.5405	0.5528	0.5643	0.5238	0.5535	0.5689
Ego.	Ungrouped	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Image–label shuffled	0.5325	0.5275	0.5625	0.5575	0.5275	0.5225
	Relation grouped	1.0000	1.0000	0.9979	0.9916	1.0000	1.0000

K Spatial Text and Written Reasoning

Text descriptions and additional reasoning offer possible ways to make heading information usable for allocentric answers. We compare the model's own image descriptions, correct heading statements supplied with images, and written reasoning. Controls distinguish the ability to repeat or follow spatial text from its use in reference-frame binding.

K.1 Answering from image descriptions and explicit heading statements

The primary analysis uses InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, and Gemma-3-4B; four additional models repeat the evaluation. Each model is tested on 148 COMFORT image pairs whose correct answers to the same binary allocentric question reverse when the reference object's heading reverses. We exclude target positions at which the geometric answer to the queried relation changes, using the reference object's coordinate system. We compare the unmodified image-and-question input with two conditions. In the self-description condition, the model first describes the image, then receives that description and the allocentric question without the image. In the explicit-heading condition, the model receives the image together with a correct statement of the reference object's heading before the allocentric question. Because descriptions could contain the answer rather than only scene information, checks include shuffling the image–description correspondence and removing descriptions containing viewpoint terms or explicit answers; descriptions that fail these checks are excluded. The paired answer response is measured by D in Eq. (1), with 0.3000 as the operational binding criterion and approximately 0.8000 as an empirical comparison for large binding effects. Separately, an open-ended task asks the model to repeat a supplied heading statement. Its repetition fidelity is the fraction of responses whose heading agrees with the supplied text.

Across eight models, heading-repetition fidelity is 0.7932–1.0000, yet both the self-description and explicit-heading conditions have $|D| \leq 0.0520$ (Table 41). The models can usually repeat the supplied heading, but neither their descriptions nor correct heading text makes answers change substantially in the direction required by heading reversal. These prompt settings do not restore reference-frame binding.

Table 41: Allocentric answers with textual scene information and a separate heading-repetition check. The first three columns report D for the original image and question, a model-generated description plus question without the image, and a correct heading statement plus image and question. The last column is the fraction of open-ended repetitions that preserve the supplied heading. Each model uses 148 pairs and 296 descriptions; results are averaged over three seeds. The first four models form the primary set, and the final four repeat the same tests.

Model	Directed correctness D			Repetition fidelity
	Baseline	Self-description	Explicit heading	
InternVL3-8B	$+0.0016$	-0.0033	-0.0279	0.9708
InternVL3.5-8B	-0.0099	-0.0039	-0.0004	0.8345
Qwen2.5-VL-7B	$+0.0090$	-0.0098	-0.0324	0.7932
Gemma-3-4B	-0.0001	-0.0251	-0.0237	1.0000
LLaVA-OneVision-7B	$+0.0113$	$+0.0031$	$+0.0124$	0.9535
llava-1.5-7B	-0.0088	-0.0145	-0.0083	1.0000
idefics2-8B	$+0.0141$	-0.0118	-0.0169	1.0000
Phi-3.5-Vision	$+0.0239$	$+0.0517$	$+0.0445$	0.9654

K.2 Answering unambiguous spatial questions from text alone

To measure use of spatial text without a coordinate transformation, we give the same four primary models an unambiguous spatial statement and a corresponding binary egocentric question, without an image. Answer accuracy checks whether the model can use the supplied spatial text to answer when no coordinate transformation is required.

All four models reach accuracy 1.0000 on these unambiguous text-only egocentric questions (Table 42).

Table 42: Text-only spatial-question control for the four primary models in Table 41. Each input contains an unambiguous spatial statement and the corresponding egocentric question, with no image. The reported value is answer accuracy averaged over three seeds.

Model	Positive-control accuracy
InternVL3-8B	1.0000
InternVL3.5-8B	1.0000
Qwen2.5-VL-7B	1.0000
Gemma-3-4B	1.0000

K.3 Effects of spatial text when the image provides conflicting evidence

We give a real image and a conflicting position statement to the model and ask a binary egocentric question. We then reverse the statement and measure the change in its answer, comparing this with the effect of reversing the same statement when no image is provided. This tests whether spatial text can influence the answer while the model also sees contradictory visual information. A separate allocentric condition supplies an incorrect heading statement without an image to test whether changing heading text changes the allocentric answer. For the egocentric comparison, the text-following ratio is

$$\mathrm{rt} = \frac{\Delta_{\mathrm{img}}}{\Delta_{\mathrm{txt}}} \in [0,\, 1], \tag{31}$$

where Δ_{img} is the answer change when the statement is reversed with an image present, and Δ_{txt} is the corresponding change without the image. A ratio near one means that the statement has a similar effect with and without visual evidence; a ratio near zero means that its effect is small when the image is present. We also add an unrelated neutral sentence and measure the fraction of answers it changes. This checks whether the effect depends on the sentence's spatial content rather than on adding any text.

Conflicting position statements yield text-following ratios of 0.3277–0.5807 for egocentric questions, whereas neutral sentences change 0.0000–0.0338 of the answers (Table 43). In the three primary models, the ratios are 0.6600–0.9200 for images on which the model has low confidence and 0.0700–0.3100 for images on which it has high confidence. In the separate allocentric test, changing heading text produces little answer response even when text is the only source of heading information. Spatial statements can influence answers in the egocentric task, including when they compete with an image, while the tested models make little use of heading statements in allocentric judgments.

Table 43: Effects of spatial and neutral statements on 148 egocentric questions per model. The text-following ratio compares answer changes with and without an image, as defined in Eq. (31). Neutral flip rate is the fraction of answers changed by an unrelated sentence. False-statement flip rate is the fraction changed toward a deliberately incorrect spatial statement. Both sentence conditions use the same question format.

Model	Text-following ratio	Neutral flip rate	False-statement flip rate
InternVL3-8B	0.3277	0.0068	0.2995
InternVL3.5-8B	0.5348	0.0338	0.5293
Qwen2.5-VL-7B	0.5807	0.0338	0.5338
Gemma-3-4B	0.5420	0.0000	0.8626

K.4 Allowing the model to write reasoning before answering

We give seven models COMFORT images with a basketball and bicycle and ask left/right or front/behind questions under both egocentric and allocentric reference frames. The four relations

are sampled uniformly over the target's azimuth relative to the camera, with the reference object facing left or right. For each question, we compare a deterministic one-word answer with a response that may contain up to 512 generated reasoning tokens before the answer. Three reasoning prompts are tested; an unrelated neutral sentence provides an additional text control. We measure balanced accuracy under each reference frame, the fraction of responses that abstain from answering, and two improvements relative to direct answering:

$$\Delta_{\mathrm{CoT}} = A^{\mathrm{allo}}_{\mathrm{CoT}} - A^{\mathrm{allo}}_{0}, \qquad R_{\mathrm{CoT}} = \frac{\Delta_{\mathrm{CoT}}}{A^{\mathrm{ego}}_{0} - A^{\mathrm{allo}}_{0}}, \tag{32}$$

Here $A^{\mathrm{allo}}_{\mathrm{CoT}}$ is balanced allocentric accuracy after allowing chain-of-thought (CoT) reasoning, and $A^{\mathrm{allo}}_{0}, A^{\mathrm{ego}}_{0}$ are balanced accuracies for direct allocentric and egocentric answers. Δ_{CoT} measures the allocentric accuracy gain. R_{CoT} expresses that gain as a fraction of the observed difference between direct egocentric and allocentric accuracy. The recovery analysis is restricted to cases with direct egocentric accuracy of at least 0.8000. Recovery then requires a reliably positive Δ_{CoT}, allocentric reasoning accuracy of at least 0.7000, and $R_{\mathrm{CoT}} \geq 0.5000$.

Written reasoning does not meet all three recovery criteria in any of the eight eligible model–axis combinations (Table 44). InternVL3-8B improves on allocentric left/right questions, but only from 0.4060 to 0.5000, while its egocentric accuracy falls. Other changes include lower allocentric accuracy in InternVL3.5-8B and increased abstention in some models; idefics2-8B's allocentric left/right accuracy falls by 0.2500. The tested prompts and reasoning lengths do not restore binding, echoing the distinction between description and spatial computation in Weng et al. (2025).

Table 44: Direct answers and answers after written reasoning for seven models on two spatial axes. L/R denotes left/right and F/B denotes front/behind. Ego. and Allo. denote egocentric and allocentric questions; all accuracy columns are balanced accuracies. Each model–axis combination has 40 direct-answer questions and 120 reasoning evaluations, comprising three prompts per question. Δ_{CoT} and R_{CoT} follow Eq. (32); Abstention is the fraction without a valid answer. The upper eight entries meet direct egocentric accuracy ≥ 0.8000 and enter the recovery analysis. The lower six describe cases without that prerequisite. Gemma-3-4B has a zero denominator for R_{CoT}, shown as n/a.

Model	Axis	Ego. accuracy		Allo. accuracy		Δ_{CoT}	R_{CoT}	Abstention
		Direct	Reasoning	Direct	Reasoning			
InternVL3-8B	L/R	0.9290	0.8510	0.4060	0.5000	+0.0940	0.1800	0.0000
InternVL3-8B	F/B	0.8570	0.7080	0.5160	0.5420	+0.0260	0.0800	0.0000
InternVL3.5-8B	L/R	0.8570	0.8690	0.5310	0.5160	−0.0160	−0.0500	0.0000
InternVL3.5-8B	F/B	0.8210	0.8510	0.5470	0.5210	−0.0260	−0.0900	0.0000
Qwen2.5-VL-7B	L/R	0.8750	0.8210	0.4060	0.4580	+0.0520	0.1100	0.0000
Qwen2.5-VL-7B	F/B	0.8570	0.7200	0.4530	0.4950	+0.0420	0.1000	0.0000
llava-1.5-7B	L/R	0.8210	0.4170	0.4690	0.3330	−0.1350	−0.3800	0.3000
Phi-3.5-Vision	L/R	0.8930	0.5930	0.4220	0.5260	+0.1040	0.2200	0.0000
Gemma-3-4B	L/R	0.5000	0.6710	0.5000	0.4640	−0.0360	n/a	0.0000
Gemma-3-4B	F/B	0.5000	0.5040	0.5000	0.5050	+0.0050	n/a	0.0000
idefics2-8B	L/R	0.6070	0.4640	0.5620	0.3120	−0.2500	−5.6000	0.3300
idefics2-8B	F/B	0.6790	0.4760	0.5780	0.4060	−0.1720	−1.7100	0.3300
llava-1.5-7B	F/B	0.6790	0.4310	0.5160	0.3330	−0.1820	−1.1200	0.2300
Phi-3.5-Vision	F/B	0.5000	0.4880	0.4690	0.5470	+0.0780	2.5000	0.0000

L Probe and Intervention Controls

Small answer effects could arise from incomplete heading removal, an ineffective intervention location, or a probe that misses available information. The controls below measure decoding after erasure, check whether the same locations support answer-changing interventions, and compare probe capacities. Equal-magnitude random perturbations and deliberately incorrect spatial text help interpret changes attributed to information content or improved binding.

L.1 Removing estimated heading directions and testing later recovery

We modify image-token residuals immediately before the layer interval in which image information affects the answer, then continue the COMFORT forward pass and measure heading decoding after that interval. The interval is the readout window identified in the preceding layer-by-layer interventions. The first analysis fits heading directions to residuals averaged over image-token positions, using InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, LLaVA-OneVision-7B, and Gemma-3-4B. The second fits directions separately at each image-token position, using seven models: InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, llava-1.5-7B, Phi-3.5-Vision, idefics2-8B, and Gemma-3-4B. In both analyses, logistic-regression directions are estimated from training representations and evaluated on separate test samples.

Let the columns of V be mutually orthogonal unit vectors spanning the estimated heading subspace. Before the readout window, each targeted residual vector x is changed to

$$x' = (I - VV^\top)\, x, \tag{33}$$

so the component in that subspace is set to zero and all orthogonal components are retained. When directions are estimated from the average image-token residual, they are fitted and removed iteratively; n_{erase} is the number removed at the reported point where decoding has substantially decreased. Because averaging can cancel information represented differently at different positions, the second analysis instead estimates and removes directions independently at every image-token position whose initial heading decoding accuracy exceeds 0.6500, without imposing a fixed dimension limit.

We first remove only the direction given by the difference between the two heading-class means. We then compare this single-direction removal with the iterative subspace removal and with removal of a random subspace containing the same number of directions. We measure heading accuracy before intervention ($\mathrm{acc}_{\mathrm{pre}}$), immediately after intervention at the window input ($\mathrm{acc}_{\mathrm{in}}$), and at the window output after the model continues computing ($\mathrm{acc}_{\mathrm{out}}$). $\mathrm{acc}_{\mathrm{rand}}$ is output accuracy after the random-subspace operation. These are probe accuracies for reference heading.

Removing the class-mean direction estimated from averaged image-token residuals leaves input heading accuracy at 1.0000 in all five models (Table 45). Removing multiple directions lowers $\mathrm{acc}_{\mathrm{in}}$ to 0.6050–0.6870, but both $\mathrm{acc}_{\mathrm{out}}$ and $\mathrm{acc}_{\mathrm{rand}}$ remain 1.0000. Input decodability is reduced rather than eliminated, and heading is fully decodable again after the window.

Table 45: Removal of heading directions estimated from averaged image-token residuals in five models. n_{erase} is the number of directions removed. $\mathrm{acc}_{\mathrm{pre}}$ is heading accuracy before removal, $\mathrm{acc}_{\mathrm{in}}$ is accuracy immediately afterward at the window input, and $\mathrm{acc}_{\mathrm{out}}$ is accuracy at the window output after continuing the model computation. $\mathrm{acc}_{\mathrm{rand}}$ instead removes a random subspace with the same number of directions. All entries are probe accuracies; post-intervention input accuracy remains above 0.5000.

Model	n_{erase}	$\mathrm{acc}_{\mathrm{pre}}$	$\mathrm{acc}_{\mathrm{in}}$	$\mathrm{acc}_{\mathrm{out}}$	$\mathrm{acc}_{\mathrm{rand}}$
LLaVA-OneVision-7B	11	1.0000	0.6870	1.0000	1.0000
InternVL3.5-8B	12	1.0000	0.6390	1.0000	1.0000
Qwen2.5-VL-7B	16	1.0000	0.6050	1.0000	1.0000
InternVL3-8B	22	1.0000	0.6420	1.0000	1.0000
Gemma-3-4B	**45**	1.0000	0.6650	1.0000	1.0000

Estimating removal directions independently at each image-token position lowers input accuracy to 0.5388–0.6253, while both window-output measures remain 1.0000 (Table 46). Even Gemma-3-4B, with 10,562 directions removed across 256 positions, retains output heading accuracy 1.0000. This count sums dimensions across positions, averaging approximately 41 per position.

In both analyses, input heading accuracy remains above 0.5000 after removal, and subsequent output accuracy reaches 1.0000. Because some heading information remains at the intervention site, we next apply a stricter linear-erasure method and measure decoding again at both sites.

Table 46: Removal of separately estimated heading directions at image-token positions in seven models. $\mathrm{acc}_{\mathrm{pre}}$, $\mathrm{acc}_{\mathrm{in}}$, $\mathrm{acc}_{\mathrm{out}}$, and $\mathrm{acc}_{\mathrm{rand}}$ have the meanings defined in Table 45. Positions counts the image-token positions whose initial heading accuracy exceeds 0.6500 and that receive the intervention. Total dimensions is the sum of removed dimensions across those positions. The random comparison removes the corresponding number of directions.

Model	$\mathrm{acc}_{\mathrm{pre}}$	$\mathrm{acc}_{\mathrm{in}}$	$\mathrm{acc}_{\mathrm{out}}$	$\mathrm{acc}_{\mathrm{rand}}$	Positions	Total dimensions
InternVL3-8B	0.9859	0.5976	1.0000	1.0000	256	1514
InternVL3.5-8B	0.9919	0.6189	1.0000	1.0000	256	2576
Qwen2.5-VL-7B	0.9836	0.6253	1.0000	1.0000	324	4908
llava-1.5-7B	0.9911	0.5691	1.0000	1.0000	576	2734
Phi-3.5-Vision	0.9935	0.6098	1.0000	1.0000	757	9661
idefics2-8B	0.9942	0.5388	1.0000	1.0000	324	2168
Gemma-3-4B	0.9876	0.6223	1.0000	1.0000	256	**10562**

L.2 Removing linearly decodable heading with a closed-form erasure method

Following amnesic probing (Elazar et al., 2021), we use LEACE (LEAst-squares Concept Erasure) (Belrose et al., 2023) to remove information that linear predictors can use to distinguish heading classes. The method equalizes class-conditional means and, under its assumptions, prevents linear predictors with convex losses from outperforming the optimal constant predictor while minimizing mean-squared changes among affine maps satisfying this constraint. We retrain heading probes after erasure and measure the effects on spatial answers separately.

The erasure checks at the readout-window input cover six models. Operators are fitted using 296 training representations, then applied separately to image-token residuals during evaluation. The behavioral comparisons that pass the positive controls cover InternVL3-8B at layer 14, InternVL3.5-8B at layer 18, and Qwen2.5-VL-7B at layer 14. Those operators use a separate set of 2,880 training representations. The training and evaluation sets are disjoint. We measure balanced heading accuracy after erasure and the relative change in the residual vector,

$$\rho = \frac{\|x' - x\|}{\|x\|}, \tag{34}$$

where x and x' are the residual vectors before and after the operation. On binary allocentric questions, the answer effect is ϵ from Eq. (5), the mean absolute change in yes probability. We also report the fraction of discrete answers that change and the effect per unit residual change, ϵ/ρ.

We compare heading erasure with three operations. The first removes randomly selected directions, using the same number as the corresponding targeted operation. The second applies LEACE to camera-azimuth sector labels at the same layer; its three modified directions are estimated from sector-conditioned representations averaged over image tokens. The third removes the entire subspace previously identified by its effect on answers in the camera-replacement experiments. This last subspace is distinct from the three directions used for sector erasure. We also scale its removal so that ρ equals that of heading erasure.

For heading erasure, an answer effect at or below both 0.0500 and the 95th percentile of the one-direction random control is treated as indistinguishable from that control. Camera-sector erasure must exceed the corresponding three-direction random control and reach 0.1000; full camera-subspace removal must exceed its random control with the same number of directions and reach 0.1000. At least one of these two camera operations must meet its criterion before a small heading effect is interpreted at that layer. The six-model window-input evaluation does not meet this behavioral positive-control requirement, so its results validate erasure and decoding only. The three-model layer-14/layer-18 evaluation does meet it through full subspace removal. Finally, output heading accuracy of at least 0.9000 is used to identify recovery of linear decodability in later layers.

At window input, heading erasure reduces heading accuracy to 0.5000 in all six models, whereas removing a random direction leaves it at least 0.9831 (Table 47). Camera-sector accuracy changes by at most 0.0068 after heading erasure. Conversely, erasing camera-sector information sharply

reduces sector decoding while heading accuracy remains 1.0000. The three layers used for behavioral evaluation show the same selective disruption: the targeted linear decoding is lost or reduced while the other tested label remains recoverable.

Table 47: Probe checks after LEACE erasure. The upper six entries intervene immediately before the readout window; the Site column gives that window's layer range. The lower three entries use the behavioral intervention layer, 14 for InternVL3-8B and Qwen2.5-VL-7B and 18 for InternVL3.5-8B. Heading decoding reports post-erasure balanced accuracy, accuracy after removing one random direction, and the 95th percentile ($q_{0.9500}$) of shuffled-label accuracy. Camera-sector columns report accuracy with unmodified residuals, after heading erasure, and after sector erasure. Heading and sector erasure modify one and three directions, respectively.

Model	Site	Heading decoding			Camera sector		
		Erased	Random rank 1	Shuffle $q_{0.9500}$	Baseline	Heading erased	Camera erased
InternVL3-8B	14–19	0.5000	1.0000	0.5350	0.8919	0.8986	0.0878
InternVL3.5-8B	18–24	0.5000	1.0000	0.5470	0.8986	0.9037	0.0878
Qwen2.5-VL-7B	14–19	0.5000	1.0000	0.5390	0.8919	0.8919	0.0878
llava-1.5-7B	11–16	0.5000	0.9831	0.5380	0.9054	0.9054	0.1014
idefics2-8B	11–16	0.5000	1.0000	0.5310	0.8716	0.8716	0.0946
Gemma-3-4B	18–24	0.5000	1.0000	0.5300	0.9054	0.9054	0.1014
InternVL3-8B	Layer 14	0.5000	1.0000	0.5160	0.9920	0.9920	0.1951
InternVL3.5-8B	Layer 18	0.5000	1.0000	0.5180	0.9780	0.9778	0.1944
Qwen2.5-VL-7B	Layer 14	0.5000	1.0000	0.5170	0.9806	0.9804	0.1958

L.3 Behavior after erasure

At layer 14 of InternVL3-8B, layer 18 of InternVL3.5-8B, and layer 14 of Qwen2.5-VL-7B, we ask binary allocentric questions before and after three interventions: LEACE heading erasure, LEACE camera-sector erasure, and complete removal of the camera-related subspace identified by earlier answer-changing interventions. We measure the same yes-probability effect ϵ for all three operations, and the fraction of discrete answers that change after complete subspace removal.

Table 48: Changes in allocentric answers after three residual interventions in the primary models. All three columns report ϵ, the mean absolute change in yes probability. Heading and Camera erasure apply LEACE to heading and camera-sector labels. Full subspace removal removes the complete camera-related subspace identified by earlier answer-changing interventions. Parentheses in the first two columns give the 95th percentiles of random controls modifying the same number of directions. Parentheses in the final column give the fraction of discrete answers that change.

Model	Heading erasure ϵ	Camera erasure ϵ	Full subspace removal ϵ
InternVL3-8B	**0.0096** (0.0225)	0.0486 (0.0282)	0.3895 (**flip 0.4090**)
InternVL3.5-8B	**0.0077** (0.0162)	0.0349 (0.0449)	0.2562 (**flip 0.2200**)
Qwen2.5-VL-7B	**0.0032** (0.0248)	0.0143 (0.0390)	0.3540 (**flip 0.4530**)

Heading erasure changes answer probabilities by only 0.0032–0.0096 across the three models, below each model's random-control threshold (Table 48). Its relative residual changes are 0.0049–0.0060, or 2.2000–10.0000 times those caused by removing one random direction. Heading accuracy returns to 1.0000 after the readout window: linear decodability is disrupted at the intervention location and restored later, with little change in the final answer.

L.4 Checking intervention sensitivity and matching residual-change magnitudes

To check whether intervention at each chosen layer can affect the answer at all, we remove the full camera-related subspace at the same layer as heading erasure. We compare this operation with random subspaces containing the same number of directions. Because removing a large subspace can cause a larger residual change than erasing heading, we also scale the full-subspace perturbation until its relative magnitude ρ matches heading erasure in the same model.

Full camera-subspace removal changes answer probabilities by 0.2562–0.3895 and flips 0.2200–0.4530 of discrete answers across the three models (Table 48). Every probability effect exceeds its random control and the 0.1000 criterion, confirming that an intervention at each selected layer can alter answers.

Complete camera-subspace removal changes residuals by 0.6599–0.8904 relative to their original norm, whereas heading erasure changes them by 0.0049–0.0060 (Table 49). After scaling the camera perturbation to the heading-erasure magnitude, its answer effects are 0.0039–0.0078 and are not consistently separated from random perturbations. Large camera-subspace removals change answers, while reducing their magnitude to match heading erasure leaves effects near the random-control values.

Table 49: Camera-subspace interventions scaled to the magnitude of heading erasure. Full-removal ρ is the relative residual change after complete removal of the camera-related subspace. Matched ρ is the change after scaling that operation and equals the heading-erasure magnitude for the same model. Matched ϵ is its effect on yes probability. Random $q_{0.9500}$ is the 95th percentile of effects from random perturbations using the same subspace dimension and matched magnitude. The definition of ρ is Eq. (34).

Model	Full-removal ρ	Matched ρ	Matched ϵ	Random $q_{0.9500}$
InternVL3-8B	0.6723	0.0049	0.0078	0.0082
InternVL3.5-8B	0.8904	0.0060	0.0039	0.0040
Qwen2.5-VL-7B	0.6599	0.0053	0.0042	0.0036

L.5 Comparing probes with different capacities on the same residuals

We use the same final-layer residual features at the answer position to compare a difference-of-means (DoM) projection, logistic regression over all residual dimensions with several regularization strengths, and probes restricted to varying numbers of principal components from principal component analysis (PCA). Training and test samples are grouped by scene, so samples from one scene never appear in both. Each probe predicts reference heading and is also evaluated with shuffled heading labels. Appendix C.2 describes the fitting and dimension scans. Because the features and divisions remain unchanged, differences in decoding accuracy here reflect the tested probe choice rather than a change in model representation.

Across eight models, DoM accuracy is 0.5203–0.8920, while full-dimensional logistic regression reaches 0.9830–1.0000 (Table 50). Five models need more than 16 principal components to reach accuracy 0.9000. Shuffled-label accuracies span 0.4089–0.5575 across all tested capacities; their within-model means across capacities are 0.4681–0.5164. DoM accuracy and the number of principal components needed for accuracy 0.9000 have rank correlation -0.9142: models with less information recoverable along the class-mean direction generally require more components.

L.6 Increasing an existing residual component along an answer-class direction

We run a single-model intervention in Gemma-3-4B at layer 17, using 166 left/right questions. A heading probe at this layer scores 0.6167, so heading-related information is detectable before intervention. We estimate a direction from the difference between the mean residuals for the two answer classes, then change the magnitude of the existing component along that direction while keeping the input question unchanged, and compare positive amplitudes, negative amplitudes, and

Table 50: Reference-heading decoding from the same final-layer answer-position residuals in eight models, with all samples from a scene assigned to the same training or test group. DoM projects onto the difference between heading-class means. Logistic uses all residual dimensions and selects among regularization strengths. PCA dimension is the first tested number of principal components that permits accuracy 0.9000. Individual shuffled-label scores span 0.4089–0.5575; their within-model means across capacities span 0.4681–0.5164. DoM accuracy and the required dimension have rank correlation -0.9142.

Model	DoM	Logistic	PCA dimension
InternVL3-8B	0.5203	0.9866	64
Qwen2.5-VL-7B	0.5233	1.0000	64
InternVL3.5-8B	0.5503	0.9830	64
LLaVA-OneVision-7B	0.5508	1.0000	32
llava-1.5-7B	0.6251	1.0000	16
idefics2-8B	0.6689	1.0000	32
Phi-3.5-Vision	0.7633	1.0000	16
Gemma-3-4B	0.8920	1.0000	8

a direction estimated after shuffling labels. This tests whether increasing that component improves answers beyond changes obtainable with a shuffled-label direction.

The model initially answers 45 questions incorrectly. For each amplitude, we report overall accuracy, the number of those errors corrected, the number of initially correct answers made incorrect, and the correction rate $N_{\text{corrected}}/45$. The stated improvement criterion requires a median correction rate of at least 0.1500 across positive amplitudes. Since a correction can be offset by a newly introduced error, correction rate is interpreted together with the introduced-error count and overall accuracy.

The median correction rate across positive amplitudes is 0.1333, below the 0.1500 improvement criterion (Table 51). At $\alpha = 2$, the estimated and shuffled-label directions each correct five of the 45 errors. Larger amplitudes correct more errors but also introduce many new ones.

Table 51: Residual-component interventions in Gemma-3-4B at layer 17, where heading-probe accuracy is 0.6167. There are 166 left/right questions and 45 baseline errors. Negative α suppresses the component along the direction estimated from the two answer-class means, and positive α increases it. Corrected counts initially incorrect answers made correct; Introduced counts initially correct answers made incorrect. Correction rate divides Corrected by 45. The final condition estimates the intervention direction after shuffling labels and applies it at $\alpha = 2$.

Setting	α	Accuracy	Corrected	Introduced	Correction rate
Baseline	0.0000	0.7289	0	0	0.0000
Suppress	-4.0000	0.6988	7	12	0.1556
Suppress	-2.0000	0.7349	2	1	0.0444
Amplify	0.5000	0.7229	1	2	0.0222
Amplify	1.0000	0.7349	6	5	0.1333
Amplify	2.0000	0.7169	5	7	0.1111
Amplify	4.0000	0.6325	18	34	0.4000
Amplify	8.0000	0.6386	12	27	0.2667
Shuffled	2.0000	0.7108	5	8	0.1111

L.7 Testing whether answers follow deliberately incorrect spatial text

We evaluate InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, and Gemma-3-4B on single-image SpinBench questions covering depth, left/right, and front/behind relations (Zhang et al., 2026). We retain questions for which correct and conflicting statements can be constructed from the question and its answer options. Both conditions retain the image and the original multiple-choice question. The

baseline adds no statement; the comparison prepends either an incorrect relation between the queried objects or a statement naming an incorrect answer option. Among questions answered correctly at baseline, we measure the fraction that become incorrect and separately check whether the new answer selects the option implied by the false statement. These statements directly convey an answer, unlike the heading-only notes used with COMFORT allocentric questions.

Adding a false spatial statement turns 0.3011–0.8869 of initially correct answers into errors, and every observed flip selects the option implied by that statement (Table 52). The answer changes follow the supplied text despite a correct baseline judgment from the image.

Table 52: Answer changes after adding deliberately incorrect spatial text to single-image SpinBench questions for four models. The image and original answer options remain in both conditions. Depth, Lateral, and Sagittal denote depth, left/right, and front/behind questions. Baseline accuracy and errors are measured without the statement. Flip rate is the fraction of initially correct answers that become incorrect; in these results, every such answer follows the false statement. Its denominator is the question count minus Baseline errors. Each model uses the same sets of 152 depth, 286 left/right, and 198 front/behind questions.

Model	Relation	Baseline accuracy	Baseline errors	Flip rate
InternVL3-8B	Depth	0.9605	6	0.3356
	Lateral	0.9755	7	0.3011
	Sagittal	0.8687	26	0.4070
InternVL3.5-8B	Depth	0.9671	5	0.7687
	Lateral	0.9301	20	0.3684
	Sagittal	0.9394	12	0.6828
Qwen2.5-VL-7B	Depth	0.9539	7	0.8690
	Lateral	0.9231	22	0.4432
	Sagittal	0.9343	13	0.4973
Gemma-3-4B	Depth	0.6382	55	0.7526
	Lateral	0.7727	65	0.8869
	Sagittal	0.5707	85	0.7257

M TRANSFER TO REAL PHOTOGRAPHS

The earlier image-pair experiments use scenes produced by a renderer, which allows the reference heading or camera viewpoint to be changed while keeping the other specified scene variables fixed. Here we examine real photographs in three tests. We train new probes on photograph representations to test whether object pose can be decoded, apply directions identified in the rendered-image experiments without fitting them again to test whether those particular directions transfer, and add the directions to model activations to test their effects on spatial answers.

M.1 MODELS AND DATA

The decoding and direction-transfer analyses cover InternVL3-8B, InternVL3.5-8B, Qwen2.5-VL-7B, llava-1.5-7B, idefics2-8B, Phi-3.5-Vision, and Gemma-3-4B. Experiments that add a direction to the residual stream focus on InternVL3-8B, InternVL3.5-8B, and Qwen2.5-VL-7B; the numerical intervention results reported in this appendix are for InternVL3.5-8B.

PASCAL3D+ (Xiang et al., 2014) provides camera-relative object azimuth annotations for real photographs. These annotations measure object pose, while the COMFORT task uses reference heading to judge the position of another object. We use its car, bus, and motorbike categories for the decoding and direction-transfer tests. The Visual Spatial Reasoning dataset (VSR) (Liu et al., 2023) provides images and yes/no spatial questions; its egocentric questions are used to measure the effect of adding a direction to the residual stream. ViewSpatial-Bench (Li et al., 2025) provides a further test of whether the intervention directions affect answers on another dataset.

Because several samples can come from the same photograph, all samples from that photograph are assigned together to either the training set or the test set when fitting and evaluating a decoder.

M.2 THREE TESTS

Training new decoders on photographs. We extract model representations of each full photograph or annotated object crop, fit a decoder on the training photographs, and evaluate it on different test photographs. One decoder predicts whether the annotated object heading is leftward or rightward; another predicts the sine of the annotated azimuth. We report classification accuracy for the first and the coefficient of determination R^2 for the second, using the explained-variance definition in Eq. (11). Decoders trained after shuffling the target labels provide controls.

Applying the same directions to photographs without fitting them again. From the analysis of rendered images, we take the difference between the mean residuals of the two heading classes, the heading direction fitted by logistic regression, the direction associated with camera-relative position, and the yes/no readout axis. We keep each direction fixed, standardize the photograph representation features, and project them onto that direction. We calculate the absolute Pearson correlations between this projection and the sine and cosine of the annotated azimuth, retaining the larger correlation. Table 54 summarizes these correlations across the three object categories using each direction's best layer. For comparison, we fit a new decoder using the same photograph dataset. Define

$$\mathrm{tr} = \frac{R^2_{\mathrm{frz}}}{R^2_{\mathrm{ret}}}, \tag{35}$$

where $R^2_{\mathrm{frz}} = r^2_{\mathrm{proj}}$ is the square of the fixed-direction correlation summarized in Table 54, and R^2_{ret} is the coefficient of determination of the decoder fitted on photographs. This ratio compares the squared correlation of the transferred direction with the measured performance of the photograph-trained decoder. A low ratio indicates poor transfer along the fixed direction relative to the decoder fitted on photographs.

Adding a direction to the residual stream. While the model answers an egocentric yes/no question about a VSR image, we add a vector along the camera-related direction at its image-token residual positions and vary the injection strength α. We measure how the yes-minus-no logit margin changes with injection strength. We repeat the operation with a random direction whose vector has the same norm and with the heading direction. On ViewSpatial multiple-choice questions, the response measure is instead $\|\boldsymbol{s}_\alpha - \boldsymbol{s}_0\|/|\alpha|$, where $\boldsymbol{s}_\alpha$ is the vector of answer-option logits after injection and $\boldsymbol{s}_0$ is the vector without injection. Within each dataset, the reported ratio divides the response to the specified direction by the response to the random direction, using the same response measure for both. We compare an initial set of 74 examples, a repeat using 74 examples, and a larger set of 2,017 examples, then test the directions on 5,712 ViewSpatial examples. The model layers and separate results for the camera and heading directions are reported in Table 55.

M.3 RESULTS

Newly fitted decoders recover object-pose information from InternVL3.5-8B's photograph representations: heading accuracy is 0.7662–0.8839, compared with 0.4937–0.5202 after label shuffling (Table 53). Sine-component regression also exceeds its shuffled baseline in all four cases, with positive R^2 for actual annotations and negative R^2 after shuffling.

Heading directions learned from rendered images transfer poorly to photographs across seven models, with transfer ratios of 0.0240–0.2560 (Table 54). Gemma-3-4B has the highest relative ratio but the weakest photograph-trained decoder ($R^2 = 0.1901$); its projection correlation is also weak in absolute terms.

For InternVL3.5-8B on VSR egocentric questions, adding the camera direction has a larger effect than adding a random direction at both sample sizes, with ratios of 4.2020–5.3160 (Table 55). The heading-direction ratio is only 1.0760 in the larger test. ViewSpatial shows a much smaller separation between camera and heading ratios (1.7560 versus 1.4400), so the measured transfer depends on the spatial task.

Table 53: Decoders trained on InternVL3.5-8B representations of PASCAL3D+ photographs. Full uses the entire photograph, and Crop uses the annotated object region. N is the number of examples. Left/right accuracy measures prediction of the annotated heading side, and sine-component R^2 measures prediction of the sine of object azimuth. Observed uses the actual annotations; Shuffled uses label-shuffled controls. Samples from one photograph remain together when separating training and testing.

Category	Input	N	Left/right accuracy		Sine-component R^2	
			Observed	Shuffled	Observed	Shuffled
Car	Full	5353	0.7674	0.5202	0.2985	−0.0330
Car	Crop	5632	0.8839	0.5181	0.6192	−0.0319
Bus	Crop	1084	0.8423	0.4937	0.4918	−0.1553
Motorbike	Crop	1257	0.7662	0.4984	0.3685	−0.1347

Table 54: Applying directions obtained from rendered images to real photographs without fitting the directions again. Heading probe is the direction fitted by logistic regression; Mean difference is the difference between the residual means of the two heading classes. Their columns report correlations with annotated azimuth, selecting each direction's best layer and averaging over the three object categories. Retrained R^2 is obtained by fitting a decoder on photographs from the same categories. Transfer ratio is the square of the Heading probe correlation divided by Retrained R^2, as in Eq. (35).

Model	Heading probe	Mean difference	Retrained R^2	Transfer ratio
InternVL3-8B	0.2026	0.1665	0.5385	0.0760
InternVL3.5-8B	0.1868	0.2420	0.4105	0.0850
Qwen2.5-VL-7B	0.2318	0.2472	0.4881	0.1100
llava-1.5-7B	0.2012	0.1927	0.5435	0.0740
idefics2-8B	0.0810	0.0905	0.2787	0.0240
Phi-3.5-Vision	0.1247	0.1273	0.3337	0.0470
Gemma-3-4B	0.2204	0.1823	0.1901	0.2560

N Extended Related Work

N.1 Linguistic and cognitive foundations

Directional terms such as left and right require a reference frame (Levinson, 1996; Frank, 1998). Levinson (2003) distinguishes absolute frames based on fixed environmental directions, intrinsic frames based on an object's own axes, and relative frames based on an observer's viewpoint, and documents cross-linguistic differences in the normally assumed frame. Human spatial judgments can use axes defined around reference objects (Logan & Sadler, 1996; Carlson-Radvansky & Logan, 1997), and communication can take the addressee's perspective into account (Danziger, 2010). In our experiments, allocentric questions require the reference object's heading to define the axes, while egocentric questions use the camera viewpoint.

N.2 Spatial evaluation and applications

Spatial benchmarks evaluate several different abilities. Kamath et al. (2023) and Wang et al. (2024) test basic directional relations between objects in natural images (Lin et al., 2014). Ma et al. (2025) also tests height, depth, occlusion, and viewpoint, while Zhang et al. (2025a) distinguishes questions about one spatial relation from questions that require several reasoning steps. Ramakrishnan et al. (2025) and Yang et al. (2025) examine sequences of views within rooms and whether spatial judgments remain consistent as the viewpoint moves. COMFORT (Zhang et al., 2025b) and SpinBench (Zhang et al., 2026) compare reference frames directly and report a preference for egocentric answers and difficulty following instructions to change the reference frame.

Table 55: Adding camera-related and heading-related directions to InternVL3.5-8B residuals while answering spatial questions. Each value compares the response to the specified direction with the response to a random direction with the same vector norm. VSR measures the change in the yes-minus-no logit margin as injection strength varies; ViewSpatial measures the answer-option logit-vector displacement per unit injection strength, $\|\boldsymbol{s}_\alpha - \boldsymbol{s}_0\|/|\alpha|$. The selected layer gives the largest separation between the camera and heading effects. VSR ego. denotes egocentric yes/no questions from VSR. Initial and Repeat use 74 examples, Expanded uses 2,017 examples, and Transfer applies the directions to 5,712 ViewSpatial examples.

Task	Run	N	Layer	Camera direction	Heading direction
VSR ego.	Initial	74	$\ell = 10$	5.3160	0.9460
VSR ego.	Repeat	74	$\ell = 10$	4.3730	0.7820
VSR ego.	Expanded	2017	$\ell = 10$	4.2020	1.0760
ViewSpatial	Transfer	5712	$\ell = 1$	1.7560	1.4400

Spatial understanding is also needed for manipulating objects, describing scenes, and annotating object relations. Tong et al. (2024) examines how visual encoders such as Contrastive Language–Image Pretraining (CLIP) (Radford et al., 2021) retain image details, and Kang et al. (2026) studies errors in position representations. Other work examines how attention is divided between image and text inputs (Chen et al., 2025), how language-based expectations affect visual question answering (Ramakrishnan et al., 2018), and why models report objects that are absent from the image (Leng et al., 2024). Our separate decoding and intervention tests ask whether heading information can be recovered from model activations and whether changing it affects the spatial answer.

N.3 Improving spatial performance

Several approaches provide intermediate descriptions of spatial structure before generating an answer. SpatialVLM (Chen et al., 2024) uses large collections of automatically generated three-dimensional spatial annotations, while Liu et al. (2025) uses depth information to structure spatial reasoning. View-on-Graph (Liu et al., 2026) and Connecting the Dots (Luo et al., 2026) use graphs of scene entities and their relations, together with agents that retrieve relevant scene information, for localization without task-specific demonstrations. Other approaches ask a model to describe an image before answering (Weng et al., 2025), generate intermediate reasoning steps (Wei et al., 2022), or reason through visual intermediate representations (Fan et al., 2025).

N.4 Mechanistic interpretability

Mechanistic interpretability asks which internal components contribute to a model's outputs and how they interact (Elhage et al., 2021). Probes train a separate predictor on model activations and compare its performance with control tasks (Hewitt & Liang, 2019). Activation and path patching replace activations from one input with activations from another, either at a selected location or along a selected computational path, to measure the effect on the output (Meng et al., 2022; Wang et al., 2023; Zhang & Nanda, 2024). Other tools represent activations through a sparse set of learned features (Huben et al., 2024) or modify activations during inference (Li et al., 2023). Work on linear representations (Park et al., 2024), decoding relations (Hernandez et al., 2024), and associating entities with their context (Feng & Steinhardt, 2024) motivates studying how directions in activation space relate to reference-frame binding. Related methods have also been applied to vision-language models (Neo et al., 2025; Jiang et al., 2025).

Spatial identifiers associate positions in an image with the words that name objects (Kang et al., 2026). Our task additionally requires expressing those positions in axes determined by a reference object's heading. A model can locate an object correctly in the image and still answer in the wrong reference frame. Studies that describe perspective changes entirely in text report related difficulties (Yang et al., 2026). Our experiments use image pairs that separately change reference heading or camera viewpoint and measure how the answer curve changes with target azimuth, directly testing how image information determines the answer.

O METRICS FOR EVALUATING REFERENCE-FRAME BINDING

This section presents the metrics from Appendices A–M as a practical checklist for evaluating reference-frame binding in future work.

For behavioral evaluation, ask the same allocentric yes/no question on heading-reversal image pairs whose correct labels are opposite, balancing questions about opposite spatial relations. Directed correctness D measures the signed probability change toward the correct answers, while sign agreement SA measures the proportion of pairs whose probability change has the correct sign (Eq. (1)). Ideal paired predictions give $D = \mathrm{SA} = 1$; high SA with small D means changes usually have the correct sign but a small average signed effect. For multiple-choice allocentric questions with unambiguous answers under both reference frames, the egocentric substitution rate s_{ego} measures the fraction of allocentric errors that equal the egocentric answer (Appendix A.3).

For evaluating the answer mapping, ask a fixed allocentric yes/no question across target azimuths under two opposite reference headings. Fit the two answer-margin curves and measure $|\Delta\varphi|$ to test whether the response pattern rotates with the reference object: the controlled setting predicts 180° for correct binding and 0° for a camera-centered response. Interpret phase only when the separately fitted mean amplitude $\bar{a} = (a_L + a_R)/2$ is sufficient, and check the measurement with synthetic responses of known phase (Appendix F).

For evaluating generalization after training, ask four-choice allocentric questions with left, right, front, and back options, crossing seen/unseen objects with trained/reversed headings. The antipodal rate p_{opp} measures predictions exactly opposite the correct direction; its random-choice baseline is 0.250. Read it alongside accuracy and per-option counts N_c in each group to distinguish successful transfer, systematic reversal, and constant predictions. This crossed error analysis separates object generalization from heading generalization. Testing unseen target azimuths and retaining performance on egocentric questions additionally checks position generalization and preservation of basic spatial ability (Appendix G).

For diagnosing reference-frame instruction processing, use images with egocentric and allocentric questions that name the same reference object and ask about the same relation, expressed through disjoint wording sets. Cross-wording transfer τ measures whether a frame-decoding direction learned with one wording set distinguishes the requested frames in another, with the direction fixed and only the decision threshold recalibrated. Shuffled-label and external-object-mention controls separate frame discrimination from wording and object-name cues (Appendix J.1).

For diagnosing information use at the answer position, use allocentric yes/no questions and their answer-position residuals. The natural-magnitude ratio r_{nat} compares answer-margin changes from reversing the heading and empirical answer components. The geometric ratio r_{probe} measures the logistic-regression heading direction's alignment with the yes/no readout axis relative to that of the empirical answer direction. Our joint diagnostic also asks identical questions on heading-reversal pairs: Δ_{beh} measures the difference in yes-answer proportions, d the standardized difference in mean answer margin, and SF the fraction of pairs with different discrete answers. Together these distinguish decodable heading, sensitivity to heading, and direct readout alignment; D is still needed to determine whether changes follow the correct geometric direction (Appendix D).

For causal diagnosis, our crossed design applies heading-reversal and camera-reversal patches under both allocentric and egocentric yes/no questions. Compare effect size ϵ across the four conditions, using heading changes under egocentric questions as a task-irrelevant control, and check residual magnitude, random perturbations, and measurement sensitivity. This comparison tests which visual information changes answers under the requested reference frame (Appendix E).

For locating information transmission and readout, use camera-reversal pairs with allocentric yes/no questions and access to intermediate activations. Joint-update recovery $\mathrm{rec} = \epsilon_{\mathrm{joint}}/\epsilon_{\mathrm{seg}}$ measures how much of the whole-residual patch effect is recovered by replacing multiple component updates. The dimension $d_{0.8000}$ measures how many channel directions recover 0.8000 of the full effect, with equal-norm controls checking content specificity. Across layers, on the log-odds scale, the fixed-subspace fraction $\mathrm{norm}(\ell)$ measures the baseline-corrected subspace effect relative to the full patch, and $\mathrm{decay}(\ell \to \ell + 1)$ compares its relative decline with the decline in available intervention range (Appendix I). At the answer position, $\mathrm{drop}(\ell)$ measures the fraction of camera sensitivity on the answer-margin scale lost when attention outputs are replaced through depth ℓ (Appendix H.4).

For testing recovery of heading information, collect residuals during allocentric yes/no answering, remove heading components before the readout window, and compare post-removal input decoding $\mathrm{acc}_{\mathrm{in}}$ with subsequent output decoding $\mathrm{acc}_{\mathrm{out}}$ and the matched random-removal control $\mathrm{acc}_{\mathrm{rand}}$. These paired measurements test whether downstream processing restores linear access to heading; ϵ separately measures the effect on answers (Appendix L).

For evaluating text-assisted methods, the text-following ratio $\mathrm{rt} = \Delta_{\mathrm{img}}/\Delta_{\mathrm{txt}}$ compares answer changes caused by reversing a spatial statement with and without an image. Use conflicting position statements with egocentric yes/no questions, complemented by heading statements with allocentric questions and a neutral-sentence control, to test content-specific text following; paired D then tests whether heading text improves binding. For reasoning methods, ask both egocentric and allocentric questions with and without reasoning. The recovery ratio $R_{\mathrm{CoT}} = \Delta_{\mathrm{CoT}}/(A_0^{\mathrm{ego}} - A_0^{\mathrm{allo}})$ expresses the allocentric accuracy gain Δ_{CoT} relative to the original frame-accuracy gap. Require adequate baseline egocentric performance and a positive, measurable denominator, and report allocentric accuracy and abstention alongside this ratio (Appendix K).