CONFLICTING DEPTH CUES AS CAUSAL PROBES OF VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can judge the same image as concave or convex, yet their final answers do not reveal how opposing interpretations are formed internally. We use Mach-card stimuli in which texture and contour support different depth directions as causal probes, independently manipulating the two visual cues to trace their links to depth representations and model decisions. We identify cue-selective and cross-cue depth-direction features in layerwise sparse autoencoders (SAEs), localize them in image space, and examine their coexistence and relative bias. We then test their functional roles through targeted steering, ablation, and cross-layer contribution replacement. In the conflict stimulus analyzed here, cue-selective features show distinct layer profiles, and candidates for opposing depth directions become jointly available in intermediate layers; the aggregate evidence nevertheless favors concavity. Targeted interventions alter the concave–convex readout. Two representative paths further indicate that **part of an earlier cue feature's effect is carried through a later depth-direction feature (the indirect, mediated component)**, with a larger effect along the concave path. These findings expose an asymmetry between candidate coexistence and decision contribution, and establish independently manipulable depth cues as a causal probe for tracing how visual evidence shapes VLM judgments.

1 INTRODUCTION

Recovering three-dimensional structure from an image requires organizing evidence that may disagree. A Mach-card fold can appear as a convex ridge or a concave valley. If texture supports convexity while contour supports concavity, an answer favoring either interpretation leaves several explanations open: competing evidence may be absent, weak, or represented without changing the readout. Option position and language priors could also shape the answer. A causal account therefore needs to compare input cues, internal candidates, and the actual output.

Spatial task scores make this distinction consequential. SpatialVLM improved spatial reasoning after training on spatial question–answer data (Chen et al., 2024), while BLINK found difficulty with relative depth and other basic visual tasks (Fu et al., 2024). Scores across these different tasks and models do not establish what evidence supported a single judgment. Cue-specific screening and interventions can locate whether a problem concerns extraction, retention of directional information, or its contribution to the output.

We used four single-cue and two congruent control classes to screen features, then evaluated them on one fixed texture-convex/contour-concave conflict image. Screening labels, conflict-image activity, and intervention effects were assessed separately. We asked where opposing candidates co-occurred, which direction their activity favored, and whether later candidates carried output changes induced at earlier layers. This was a test of local effects on one target, not a factorial manipulation of cue strength (Figure 1).

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.

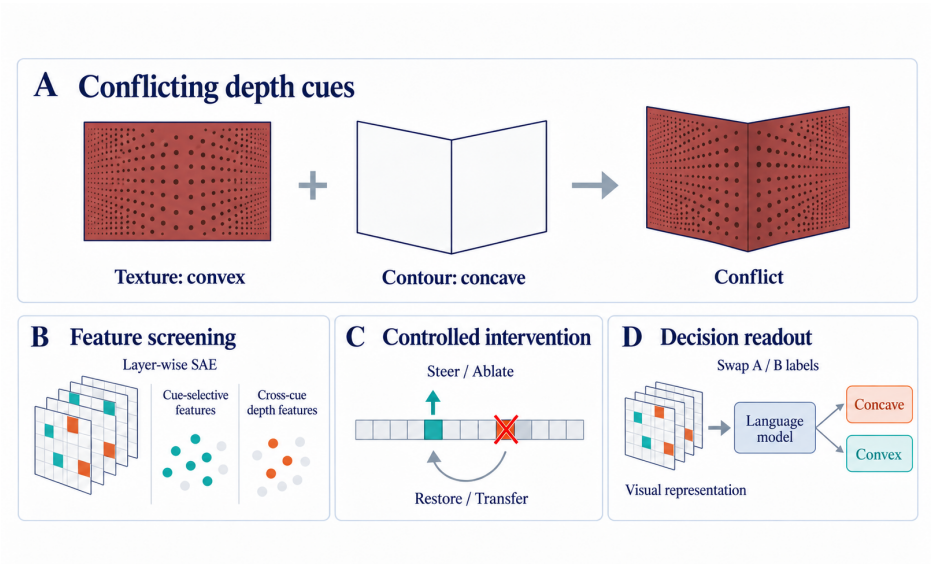

Figure 1: Causal probing framework. Single-cue and congruent controls defined SAE feature sets. Their layerwise availability and spatial responses were measured on the fixed conflict image, followed by steering, ablation, and restoration/transfer tests. The answer mapping was swapped when computing the output score. The figure illustrates the design, not a tested grid of cue strengths.

2 Related Work

2.1 From spatial performance to the use of visual evidence

Depth cues must be distinguished from the judgments they support: ordinal depth concerns relative order and metric depth distance. BLINK, GeoMeter, 3D-PC, and SPACE exposed distinct spatial demands hidden by aggregate scores (Fu et al., 2024; Azad et al., 2025; Linsley et al., 2025; Ramakrishnan et al., 2025). SpatialVLM, SpatialRGPT, and DepthLM improved tested spatial tasks (Chen et al., 2024; Cheng et al., 2024; Yeh et al., 2026). Yet improved scores cannot identify the evidence used on an individual image. Depth decodable from a representation is accessible to a probe; whether it contributes to the model's own judgment requires intervention.

2.2 Cue conflict and ambiguous images

Ambiguous images permit alternative interpretations but do not isolate their visual causes. Shape–texture conflict revealed texture dependence in image classifiers (Geirhos et al., 2018). Visual-illusion tasks exposed incomplete agreement between VLMs and human judgments (Zhang et al., 2023); naming a familiar illusion may draw on category knowledge, whereas judging its depicted geometry requires evaluating the image. Bistable-image tests further varied transformations and prompts (Panagopoulou et al., 2024). Explicitly specifying two competing image cues gives the ambiguity a controlled visual source.

Agate (2026) observed simultaneous activation of opposing feature pools in 50 of 69 bistable images with both pools available, yet steering altered captions only for some stimulus families. O3-D instead manipulated pictorial depth cues to separate behavioral cue effects from language bias (Liu et al., 2026b). Together these results leave a concrete link open: whether a specified input cue influences the answer through a selected internal candidate. We probed that link with control-based screening followed by interventions on a fixed conflict image.

2.3 Sparse features as intervention targets

SAEs decompose activations into coordinates for interpretation and intervention (Huben et al., 2024; Gao et al., 2025). PatchSAE localized visual concepts to image patches (Lim et al., 2025), but

Table 1: Operational screening criteria. T^- and C^+ were controls for the focal T^+/C^- conflict; their cue-selective counts were not reported. The cross-cue criterion requires directional agreement across cue families and should be distinguished from the pure-condition criterion.

Set	Operational comparison	Interpretation
T^+ / C^-	Pure target class vs other five classes: AUROC > 0.990; other-cue direction AUROC $\in$ [0.450, 0.550]	Pure-cue condition preference
E^+ / E^-	Same directional relation across texture, contour, and congruent comparisons	Direction consistency across cue families

coordinate identity need not be unique or stable across training runs (Leask et al., 2025). Evaluation of interpretability and control must be tied to the task computation being tested (Makelov et al., 2025). We therefore call an SAE coordinate a feature and a coordinate passing the cue screen a candidate; its label alone does not establish a causal role.

3 METHODS

3.1 STIMULI AND SCOPE OF THE ANALYSIS

We used three stimulus roles. Single-cue controls comprised convex texture (T^+), concave texture (T^-), convex contour (C^+), and concave contour (C^-), with 25 images per class. Texture controls retained a dot gradient on a neutral outline; contour controls varied the projected outline without directional texture. Two congruent classes, each with 25 images, combined cues pointing in the same direction, giving 150 controls in total. The intervention target was one T^+/C^- conflict image. The mirrored T^-/C^+ conflict was not tested. Congruent controls contributed to screening and were not assumed behaviorally equivalent to single-cue images.

The stimulus roles allowed us to test whether a coordinate distinguished a pure-cue condition, retained direction selectivity across cue families, and affected the conflict target. We did not estimate cue-strength interactions or a benefit of congruence. Texture-gradient studies motivated the texture manipulation (Higashiyama & Yamazaki, 2016). Preston et al. (2009) manipulated Mach-card shape through binocular disparity; our monocular contour was the projected folded-card outline. Shared geometry provides a stimulus analogy, not equivalent depth manipulations.

3.2 MODEL AND FEATURE SCREENING

We analyzed LLaVA-1.6-Vicuna-7B with its CLIP ViT-L/14-336 vision encoder (24 layers, hidden dimension 1,024, and 576 patch tokens after removing the class token). We trained an independent SAE at each layer using a dictionary of 65,536 features, Top-K sparsity with K = 32, and unit-normalized decoder directions. CC3M supplied 200,000 training image–text examples and a separate 10,000-image validation split. AdamW used a learning rate of 3×10^{-4}, zero weight decay, and an effective patch batch size of 1,024. Layerwise analyses used 56,250-step checkpoints; the L22 causal analysis used the 50,000-step checkpoint (Appendix A). An SAE represents an activation as $h = Wz + b + r$, where r is the reconstruction residual. Layer labels L0–L23 index encoder depth, not processing time.

We screened coordinates active on the six control classes. For T^+, 25 pure texture-convex positives were contrasted with the other 125 controls, including congruent-convex images; candidates required one-versus-rest AUROC > 0.990 and C^+-versus-C^- AUROC within [0.450, 0.550]. C^- used the analogous criterion with T^+-versus-T^- discrimination near chance. This screen identifies pure-condition preference and can omit a feature also active for a congruent image with the same cue. Near-chance discrimination of the other cue does not prove cue independence.

A separate screen identified cross-cue convex (E^+) or concave (E^-) candidates with the same directional relation across texture, contour, and congruent comparisons (Table 1). T^+/C^- therefore labels tested condition preference; E^+/E^- labels tested cross-cue direction. These control-derived labels preceded intervention on the conflict image. The exact cross-cue numerical thresholds were not supplied in the available record, limiting independent reconstruction of this screen.

3.3 Feature availability and normalized evidence

We counted screened features at each layer and mapped their patch activations to the conflict image. Top-1 used the highest-ranked responding feature; Top-3 used at most three. For example, a set with two candidates contributed both to Top-3, whereas a singleton gave identical Top-1 and Top-3 membership. Empty sets remained missing. Spatial maps described where selected features responded, without providing an independent localization-accuracy test.

To compare activity on the target against controls, we standardized each feature's image-level activity a using its control mean μ and sample standard deviation σ. For a feature set S, we averaged standardized scores before applying the sigmoid:

$$q_f = \frac{a_f - \mu_f}{\sigma_f + 10^{-6}}, \qquad E_S = \frac{1}{|S|} \sum_{f \in S} q_f, \qquad P_S = \frac{1}{1 + e^{-E_S}}. \tag{1}$$

The additive stabilizer prevents division by zero. P is normalized feature evidence, not a probability of perception. Duality is the smaller of the two direction scores, and Bias is convex evidence minus concave evidence:

$$\text{Duality} = \min(P_{\text{convex}}, P_{\text{concave}}), \quad \text{Bias} = P_{\text{convex}} - P_{\text{concave}}. \tag{2}$$

For cue comparisons, the two scores came from T^+ and C^-; for cross-cue comparisons, they came from E^+ and E^-. Bias < 0 favors concavity. Duality = 0.500 is a descriptive reference: both direction scores must reach 0.500 for Duality to reach it. It is not a significance threshold. Full-set and Top-3 summaries tested how strongly the aggregation depended on a few features (Appendix A).

3.4 Output readout and interventions

We computed semantic margins using two A/B answer mappings. For the concave-versus-convex wording, each margin was the concave-option logit minus the convex-option logit; their mean M used a common semantic sign. A second baseline wording contrasted a convex roof ridge with a concave wall corner and is reported separately. For intervention aimed at direction d, the unmodified target defined M_0 and we used:

$$M = \frac{m_1 + m_2}{2}, \qquad \Delta M_d = s(M_{\text{intervention}} - M_0). \tag{3}$$

Here s = +1 for concavity and -1 for convexity. Positive ΔM moves the logit score toward the named interpretation; negative ΔM moves it away. This change is neither a choice probability nor evidence of a reversed answer without knowing whether the absolute margin crossed zero. A/B averaging does not establish agreement between the two mappings, which we considered separately.

Steering increased the selected contribution along its decoder direction. The main strengths were α = 0, 0.500, 1, 2, and 4, with 8 and 16 included in the supplementary scan. Ablation removed a fraction $\beta \in$ 0, 0.250, 0.500, 0.750, 1 of the current contribution; β = 1 removed it fully. Other coordinates and the reconstruction residual were preserved. Steering tested whether increasing a selected contribution could move the readout; ablation tested how the contribution already present affected the current state. Zero-dose operations defined no-change controls. Neither operation by itself implies that the named feature is a complete depth representation.

3.5 Cross-layer contribution replacement

We tested whether output changes induced at an earlier layer were carried partly by selected features at a later layer. U denotes an earlier cue-selective set and R a later cross-cue direction set. The "upstream effect" is the output change after steering U alone. The "contribution change" at R is the change in the decoded contribution of its selected SAE coordinates. We compared the baseline, steering U alone, steering U while restoring R to its baseline contribution, and transferring the R change into the unmodified run. Restoration and transfer were defined by:

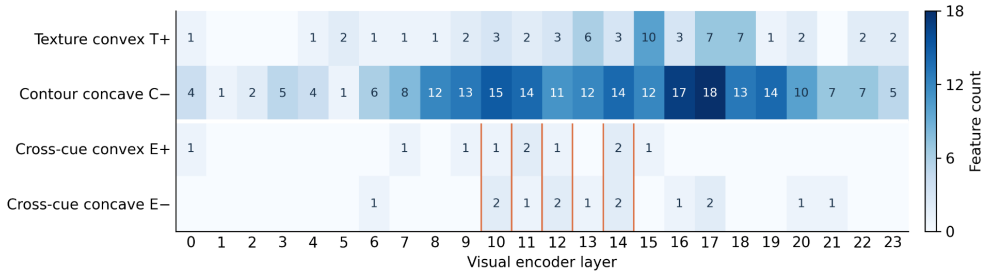

Figure 2: Screened feature counts by layer. The upper rows show the focal T^+ and C^- sets; the lower rows show E^+ and E^-. Blank cells indicate that no coordinate passed the screen. Orange boundaries mark the layers with both direction sets. T^- and C^+ counts were not part of this reported target analysis; their omission does not imply absence of the corresponding features.

$$\mathbf{h}_R^{\text{res}} = \mathbf{h}_R^U + \sum_{f \in R} (z_f^0 - z_f^U)\mathbf{W}_f, \tag{4}$$

$$\mathbf{h}_R^{\text{tr}} = \mathbf{h}_R^0 + \sum_{f \in R} (z_f^U - z_f^0)\mathbf{W}_f. \tag{5}$$

Superscripts 0 and U denote baseline and U-steered states; W_f denotes the decoder direction of feature f. All unselected contributions and the residual were retained from the state being modified. If restoration reduced the U effect and transfer alone changed the output in the same direction, R carried a separable component of that effect. This test concerns a selected contribution, rather than proving a unique or complete mediating pathway. Same-direction pairs linked C^- to later E^- or T^+ to later E^+. Appendix B reports the full scan and all qualifying pairs at $\alpha = 4.000$.

4 Results

4.1 Baseline choices showed answer-position dependence

We evaluated the conflict target using two phrasings and two swapped A/B mappings per phrasing. In all four forward passes, the chosen option occupied position A; its logit exceeded B by 1.063–1.328. After semantic reorientation and averaging, the concave-minus-convex margin was +0.211 for the shape wording. For the second wording, the roof-ridge-minus-wall-corner margin was -0.281, also on the concavity side after sign interpretation. Because each mapping selected A, neither average demonstrates a mapping-stable semantic choice. These were single greedy decisions rather than response frequencies. Subsequent target-directed intervention changes refer to logit margins; without per-mapping post-intervention logits, they cannot establish answer reversals. This position effect is essential when interpreting Tables 2 and B2.

4.2 Cue selectivity was widespread but joint direction availability was localized

Concave-contour candidates occurred at all 24 layers (1–18 per layer), reaching 17–18 at L16–L17. Convex-texture candidates occurred at 20 layers and were absent at L1–L3 and L21; 16 of their 20 layers had 1–3 candidates, with a maximum of 10 at L15. Cross-cue convex candidates first occurred at L0 and concave candidates at L6. Both directions were present together only at L10, L11, L12, and L14 (Figure 2), with at most two candidates in each direction set. Thus, early detection of one direction did not establish joint availability. These layer numbers locate screened features within this encoder; they do not mark universal functional stages.

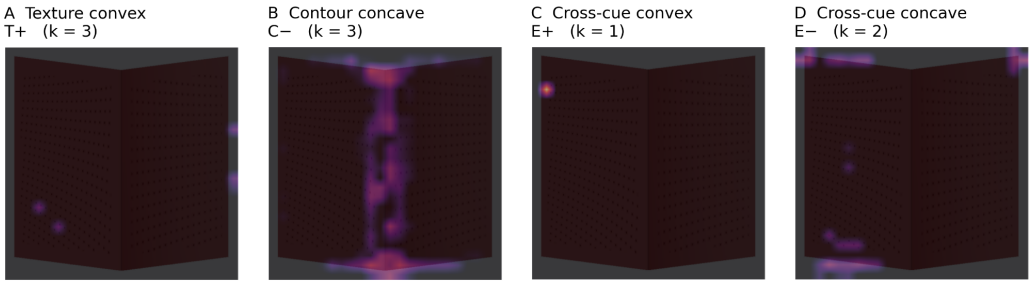

Figure 3: Spatial responses at L10. Actual Top-3 set sizes were 3, 3, 1, and 2. T^+: f58523, f11975, f11852; C^-: f51034, f13343, f56130; E^+: f27210; E^-: f49688, f33382. Colors describe within-image spatial distributions and do not support absolute magnitude comparisons across sets.

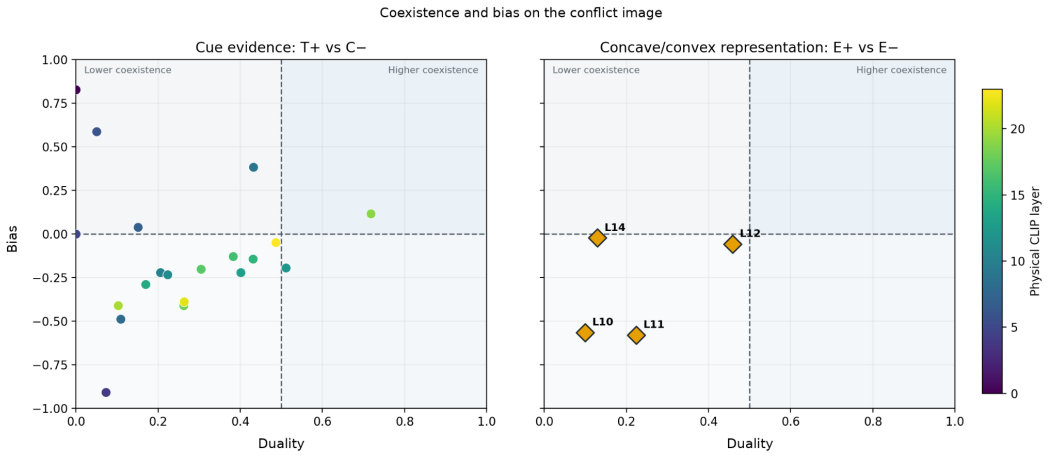

Figure 4: Coexistence and bias on the fixed conflict image. Left: T^+ versus C^-; right: E^+ versus E^-. Duality = min(Pconvex, Pconcave); Bias = Pconvex - Pconcave. Diamonds identify L10, L11, L12, and L14. Dashed lines indicate Duality = 0.500 and Bias = 0.000.

4.3 Spatial responses and evidence bias distinguished the candidates

We inspected L10 because it was the first layer at which both cross-cue direction sets were available. Texture-convex responses occupied a few texture regions, while contour-concave responses followed the central fold and upper and lower boundaries (Figure 3). E^+ responded mainly near a left edge; E^- covered upper-left and lower-left regions and additional local patches. Cross-cue selectivity therefore remained compatible with responses to local image structure. These maps described spatial support, not complete three-dimensional scene reconstructions.

All four layers with both directions available had negative Bias (Figure 4). At L12, Duality = 0.459 and Bias = -0.058 implied normalized convex and concave scores of approximately 0.459 and 0.517. At L10 these scores were 0.099 and 0.665; at L11 they were 0.224 and 0.804, reconstructed from the reported metrics. L12 therefore had relatively balanced scores, whereas the other two layers retained weaker convex evidence beside stronger concave evidence. These summaries are standardized activations, not choice probabilities; co-occurrence alone does not imply balanced evidence.

4.4 Interventions changed the output score in both directions

Table 2 groups five illustrative interventions by role. C^- L11 and T^+ L12 steering supplied the earlier interventions for the two displayed paths; C^- L02 showed removal of an existing cue contribution; E^+ L07 and E^+ L10 tested direct interventions on cross-cue candidates. This was a retrospective display from the scan, not a ranking of layers. C^- L11 steering shifted the concavity-

Table 2: Illustrative single-set interventions. Dose denotes α for steering and β for ablation; all changes are relative to the unmodified target. Positive values favor the named direction. The E^+ values were +0.0128 and -0.0262 before rounding. Top-3 always denotes at most three available features. These five rows do not establish which layer has the largest effect across the full scan.

Candidate / layer	**Operation**	**Dose**	**Selection**	**Target-directed change**
Concave contour C^- L11	Steering	4.000	Top-3	+0.172
Convex texture T^+ L12	Steering	4.000	Top-3	+0.023
Concave contour C^- L02	Ablation	1.000	Top-1	-0.320
Cross-cue convex E^+ L07	Steering	4.000	Top-1	+0.013
Cross-cue convex E^+ L10	Ablation	1.000	Top-1	-0.026

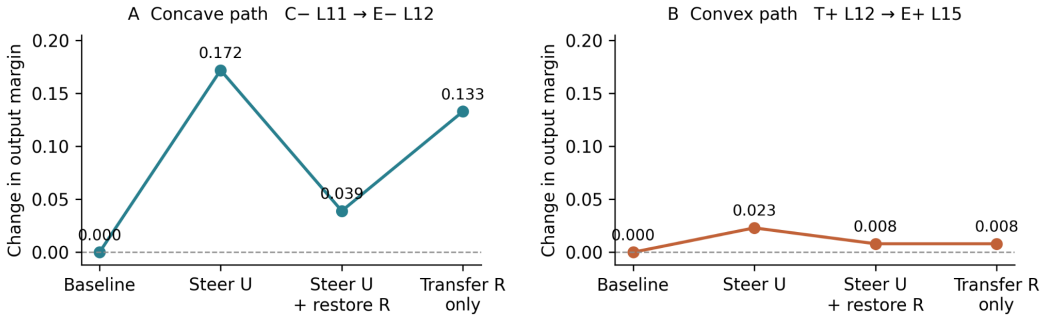

Figure 5: Four-condition tests at α = 4.000. Both panels use the same vertical scale. "Steer U + restore R" restores the selected later contribution to baseline after earlier steering; "Transfer R only" inserts the later change into an unmodified run. Target-directed change is ΔMd = s(Mintervention - M_0), as defined in Section 3.4. Values are rounded to three decimals; baseline 0.000 denotes zero change, not a neutral absolute judgment.

directed margin by +0.172 and T^+ L12 steering shifted the convexity-directed margin by +0.023. C^- L02 ablation yielded -0.320 in the concavity direction. The two cross-cue changes were +0.0128 and -0.0262 at source precision. Their nonzero values must not be confused with zero-dose controls, and small changes are sensitive to the numeric scale (Appendix B).

The intervention signs linked screened candidates to output changes, but feature labels did not determine the sign in every target state. For example, ablating an L14 convex set increased the convexity-directed score in supplementary conditions (Appendix B.3). We therefore evaluated the observed change for each operation and state rather than treating a convex screening label as sufficient evidence for a positive local contribution.

4.5 Later candidates carried part of earlier intervention effects

Among 696 earlier-cue/later-direction pairs, 6,960 nonzero dose conditions were scanned and 200 met three joint criteria for candidate change, output change, and consistent contribution direction. At α = 4.000, 15 same-direction paths qualified: 12 concave and three convex (Appendix B). We displayed C^- L11 $\rightarrow$ E^- L12 for the largest transfer change among the concave paths and T^+ L12 $\rightarrow$ E^+ L15 for the largest upstream steering change among the convex paths. Both used Top-3 and α = 4.000, the largest primary steering dose. These selection rules were retrospective and did not supply independent confirmation.

For the concave path, U steering changed the margin by 0.172; restoring R reduced it to 0.039, a reduction of 0.133. Transferring R alone produced 0.133. These values were consistent with R carrying a substantial component of that particular U effect. For the convex path, the corresponding values were 0.023, 0.008, and 0.008. Restoration reduced the U effect, but the transferred change did not equal the removed change (0.015 from the rounded values). We therefore did not assume an additive decomposition or interpret these ratios as mediated proportions. The two paths differed in

magnitude within this target, rather than establishing a general asymmetry between all concave and convex computations.

All path effects used the mapping-averaged score. None of the 15 same-direction paths had the same change direction in both mappings. Further, 138 of the 200 qualifying dose conditions had an absolute earlier-intervention effect no larger than four typical FP16 margin steps (approximately 0.031; Appendix B). These selected dependencies were sensitive to answer mapping and numerical scale; they did not establish reliable alternation between interpretations.

5 Discussion

The findings separated three questions: was a candidate selected on cue controls, how strongly did it respond to the conflict image, and did an intervention change the output score? Opposing candidates co-occurred at intermediate layers, but their normalized evidence favored concavity. Selected changes to cue contributions moved the output margin in both directions, and later contributions carried parts of two earlier effects. Crucially, the baseline choice was governed by answer position across the tested mappings. The local cue-to-candidate-to-logit links therefore constrain internal processing without establishing a stable semantic depth judgment (Agate, 2026; Liu et al., 2026b).

Candidate coexistence did not entail equal evidence or perceptual alternation. Human cue-integration accounts permit unequal weights (Landy et al., 1995; Ernst & Banks, 2002), but the present feature scores did not estimate human weights. The position effect made the gap between internal activity and a stable answer particularly visible: a normalized concavity bias coexisted with four position-dominated choices. This pattern motivates tests of how visual evidence, language wording, and answer mapping interact at the readout.

The study analyzed one model and one T^{+}/C^{-} conflict image. It did not test the reverse conflict or cue-strength interactions. SAE quality varied by layer and feature identity was not assessed across training seeds (Appendix A; Leask et al., 2025; Song et al., 2026). Generality requires a texture $\times$ contour grid, matched random interventions, mapping-specific absolute logits, and replication across stimuli. Such experiments can test whether local candidate links persist when both the visual evidence and readout change.

6 Conclusion

Conflicting depth cues exposed a separation between available candidates, their activity bias, and their local influence on output scores. We observed joint availability of opposing features, concavity-favoring normalized evidence, and candidate-specific intervention effects. Yet all reported baseline choices favored option A after swapping the mappings, and no selected cross-layer path was directionally stable across mappings. The result is a bounded mechanistic probe of cue contributions, with stable semantic depth judgments remaining to be demonstrated.

REPRODUCIBILITY STATEMENT

Section 3 specifies the control classes, model and SAE configuration, feature-screening logic, evidence metrics, and intervention comparisons. Appendix A reports layerwise SAE quality; Appendix B lists full-scan counts and all 15 same-direction paths at $\alpha = 4.000$; Appendix C specifies the interpretation of answer mappings and numerical precision. Figures 2–4 preserve the reported counts or spatial/plot coordinates, and Figure 5 uses the tabulated four-condition values on a common scale. Reproduction requires the stimulus-generation parameters, screening implementation, layerwise checkpoints, exact prompts, and separate logits for both answer mappings. The available baseline records supplied position-conditioned choices and two averaged margins; they did not include the four individual logits or intervention-specific absolute margins. Exact cross-cue thresholds and full screening code are still needed for independent reconstruction.

ETHICS STATEMENT

This study analyzed model responses to procedurally generated images and involved no new human-participant experiments. Model judgments of concavity or convexity and sparse features should not be interpreted as direct evidence of subjective experience. The study was intended to diagnose the encoding and use of visual evidence; its interpretations are limited to the analyzed stimuli, model, and intervention conditions.

AI USE STATEMENT

Generative AI assisted with literature search, manuscript editing, translation between English and Chinese, figure-label revision, and document formatting. Image generation was used only for the schematic research framework. Experimental observations were not generated by AI; numerical findings were taken from the research materials, and explicitly identified arithmetic summaries were computed from reported values. The authors are responsible for reviewing the scientific claims, source records, and final submission.

REFERENCES

Arohan Agate. Vision-Language Asymmetry in Bistable Image Captioning. *arXiv preprint arXiv:2606.08031*, 2026.

S. Aida, S. Fukamachi, and K. Shimono. Depth perception of stereoscopic transparent stimuli with frame manipulation, 2024. Scientific Reports, 14, 6712. doi:10.1038/s41598-024-57283-9.

D. Au, J. Tong, R. S. Allison, and L. M. Wilcox. Conflicting ordinal and metric depth information interferes with depth matching, 2026. Journal of Vision, 26(5), 11. doi:10.1167/jov.26.5.11.

Shehreen Azad, Yash Jain, Rishit Garg, Vibhav Vineet, and Yogesh Rawat. Understanding depth and height perception in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3650–3659, 2025.

N. Budny, K. Ghods, D. Campbell, R. Marjieh, A. Joshi, S. Kumar, J. D. Cohen, T. W. Webb, and T. L. Griffiths. Visual Serial Processing Deficits Explain Divergences in Human and VLM Reasoning, 2025. arXiv preprint. arXiv:2509.25142.

Z. Cai, Z. Liu, Y. Xiong, Z. Liu, V. Chandra, and Y. Shi. VLM³: Vision-Language Models Are Native 3D Learners, 2026. arXiv preprint. arXiv:2605.30561.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

W. Chow, J. Mao, B. Li, D. Seita, V. Guizilini, and Y. Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding, 2025. ICLR. arXiv:2501.16411.

J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth, 1995. In W. Epstein & S. Rogers (Eds.), Perception of Space and Motion, pp. 69–117. Academic Press.

D. Eigen, C. Puhrsch, and R. Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, 2014. NeurIPS. arXiv:1406.2283.

M. El Banani, A. Raj, K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3D Awareness of Visual Foundation Models, 2024. CVPR, pp. 21795–21806. arXiv:2404.08636.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Z. Fan, J. Zhang, R. Li, J. Zhang, R. Chen, H. Hu, K. Wang, H. Qu, D. Wang, Z. Yan, H. Xu, J. Theiss, T. Chen, J. Li, Z. Tu, Z. Wang, and R. Ranjan. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction, 2026. CVPR. arXiv:2505.20279.

J. M. Foley. Binocular distance perception, 1980. Psychological Review, 87(5), 411–434.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

J. J. Gibson. The Ecological Approach to Visual Perception, 1979. Houghton Mifflin.

A. Hackett, A. Denis-Remillard, and A. Cassou. From Recovery to Drop-off: How Action Post-training Reduces a VLM's Late-Layer Depth Decodability, 2026. arXiv preprint. arXiv:2608.08904.

G. Harding and J. M. Harris. Learning to Use Illumination Gradients as an Unambiguous Cue to Three Dimensional Shape, 2012. PLOS ONE, 7, e35950. doi:10.1371/journal.pone.0035950.

R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision, 2004. 2nd ed. Cambridge University Press.

Atsuki Higashiyama and Tadashi Yamazaki. Anisotropic perception of slant from texture gradient: Size contrast hypothesis. *Attention, Perception, & Psychophysics*, 78(2):647–662, 2016.

Y. Hu, V. DuTell, A. R. Girshick, and J. E. Corbett. A Framework for Generating Semantically Ambiguous Images to Probe Human and Machine Perception, 2026. arXiv preprint. arXiv:2603.24730.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

S. Joseph, P. Suresh, E. Goldfarb, L. Hufe, Y. Gandelsman, R. Graham, D. Bzdok, W. Samek, and B. A. Richards. Steering CLIP's Vision Transformer with Sparse Autoencoders, 2025. CVPR Workshop on Mechanistic Interpretability for Vision. arXiv:2504.08729.

Michael S Landy, Laurence T Maloney, Elizabeth B Johnston, and Mark Young. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision research*, 35(3):389–412, 1995.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, volume 2025, pp. 53617–53642, 2025.

B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. SEED-Bench: Benchmarking Multimodal Large Language Models, 2024. CVPR, pp. 13299–13308.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 24444–24469, 2025.

Drew Linsley, Peisen Zhou, Alekh Ashok, Akash Nagaraj, Gaurav Suhas Gaonkar, Francis Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. In *International Conference on Learning Representations*, volume 2025, pp. 30472–30494, 2025.

X. Liu, M. Luo, A. Chatterjee, H. Wei, C. Baral, and Y. Yang. Investigating VLM Hallucination from a Cognitive Psychology Perspective: A First Step Toward Interpretation with Intriguing Observations, 2025a. arXiv preprint. arXiv:2507.03123.

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin. MMBench: Is Your Multi-modal Model an All-around Player? ECCV, 2024. arXiv:2307.06281.

Y. Liu, M. Ma, X. Yu, P. Ding, H. Zhao, M. Sun, S. Huang, and D. Wang. SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning, 2025b. NeurIPS. arXiv:2505.12448.

Y. Liu, F. Zhan, K. Zhou, Y. Du, P. P. Liang, and H. Pfister. Abstract 3D Perception for Spatial Intelligence in Vision-Language Models, 2026a. CVPR, pp. 38647–38656. arXiv:2511.10946.

Yiqian Liu, Iuliia Kotseruba, and John K Tsotsos. Disentangling Pictorial Cue Understanding from Language Bias in VLMs via Depth Ordering Task. In *European Conference on Computer Vision*, pp. 643–660. Springer, 2026b.

A. Makelov, G. Lange, A. Geiger, and N. Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching, 2024. ICLR.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, volume 2025, pp. 33588–33636, 2025.

Y. Man, S. Zheng, Z. Bao, M. Hebert, L. Gui, and Y. Wang. Lexicon3D: Probing Visual Foundation Models for Complex 3D Scene Understanding, 2024. NeurIPS. arXiv:2409.03757.

D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, 1982. W. H. Freeman.

S. E. Palmer. Vision Science: Photons to Phenomenology, 1999. MIT Press.

Artemis Panagopoulou, Coby Melkin, and Chris Callison-Burch. Evaluating vision-language models on bistable images. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 8–29, 2024.

F. Peveri, F. Barban, A. Canessa, and S. P. Sabatini. Depth perception changes following adaptation to cue-dependent invariants, 2025. Scientific Reports, 15, 45741. doi:10.1038/s41598-025-28193-1.

Tim J Preston, Zoe Kourtzi, and Andrew E Welchman. Adaptive estimation of three-dimensional structure in the human brain. *The Journal of Neuroscience*, 29(6):1688–1698, 2009.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, volume 2025, pp. 101439–101498, 2025.

L. M. Schulze Buschoff, E. Akata, M. Bethge, and E. Schulz. Visual cognition in multimodal large language models, 2025. Nature Machine Intelligence, 7, 96–106. doi:10.1038/s42256-024-00963-y.

W. Shi, S. Li, T. Liang, M. Wan, G. Ma, X. Wang, and X. He. Route Sparse Autoencoder to Interpret Large Language Models, 2025. EMNLP, pp. 6801–6815. doi:10.18653/v1/2025.emnlp-main.346.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona Diab, Virginia Smith, and Kun Zhang. Mechanistic Interpretability Should Prioritize Feature Consistency in Sparse Autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210, 2026.

G. Tangtartharakul and K. R. Storrs. Visual language models show widespread visual deficits on neuropsychological tests, 2026. Nature Machine Intelligence, 8, 209–219. doi:10.1038/s42256-026-01179-y.

S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, 2024. CVPR, pp. 9568–9578. arXiv:2401.06209.

S. Ullman. The interpretation of structure from motion, 1979. Proceedings of the Royal Society B, 203, 405–426.

Q. Vo, P. Dao, C. Nguyen, and T. Nguyen. When Depth Is Better Told Than Shown: Depth-Ordinal Prompting for Vision-Language Spatial Reasoning, 2026. arXiv preprint. arXiv:2607.11173.

B. G. Wundari, I. Fujita, and H. Ban. Human and artificial visual systems share a computational principle for transforming binocular disparity into depth representation, 2025. Communications Biology, 8, 1042. doi:10.1038/s42003-025-08474-1.

Z. Xu, J. Pang, A. Anzai, and G. C. DeAngelis. Flexible computation of object motion and depth based on viewing geometry inferred from optic flow, 2026. Nature Communications, 17, 1092. doi:10.1038/s41467-025-67857-4.

C. Yao, Z. Liu, J. Zeng, L. Yu, Y. Wu, and Y. Jia. 3D Visual Illusion Depth Estimation, 2025. NeurIPS. arXiv:2505.13061.

Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, Yangyang Shi, et al. Depthlm: Metric depth from vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 134491–134505, 2026.

H. Yu, X. Qu, Y. Wang, J. Zhu, and L. Ke. Unlocking Dense Metric Depth Estimation in VLMs, 2026. arXiv preprint. arXiv:2605.15876.

T. Yuan, Y. Liu, C. Lu, Z. Chen, T. Jiang, and H. Zhao. DepthVLA: Enhancing Vision-Language-Action Models with Depth-Aware Spatial Reasoning, 2025. arXiv preprint. arXiv:2510.13375.

G. Zhan, C. Zheng, W. Xie, and A. Zisserman. A General Protocol to Probe Large Vision Models for 3D Physical Understanding, 2024. NeurIPS. arXiv:2310.06836.

Y. Zhang, Z. Zhang, X. Wei, X. Liu, G. Zhai, and X. Min. IllusionBench+: A Large-scale and Comprehensive Benchmark for Visual Illusion Understanding in Vision-Language Models, 2025. arXiv preprint. arXiv:2501.00848.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 5718–5728, 2023.

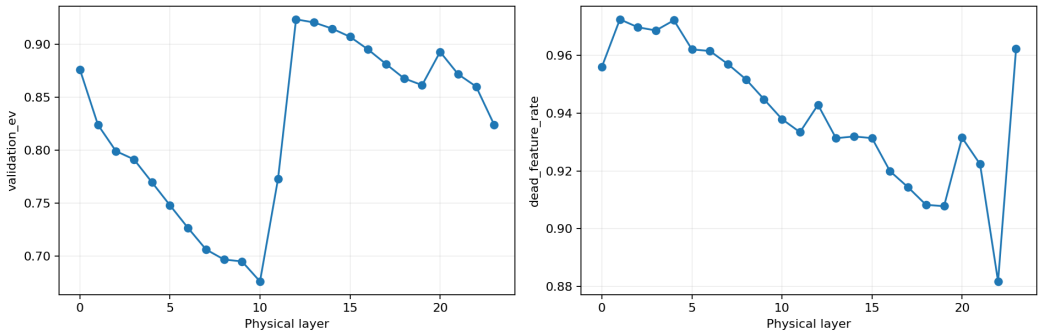

Figure A1: SAE quality by layer. Left: validation explained variance. Right: fraction of inactive features. Curves correspond to the layerwise checkpoints used in this study and are not averages across random seeds. Original result coordinates are retained.

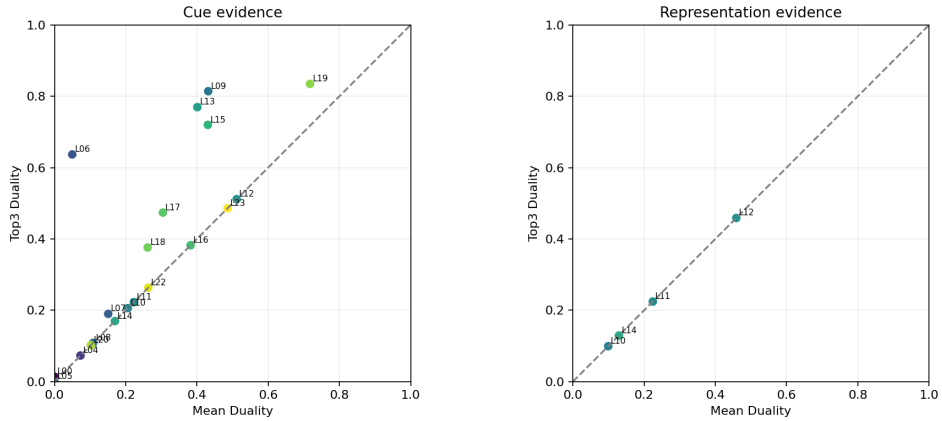

Figure A2: Comparison of coexistence metrics from full-set means and Top-3 features. The dashed line is y = x. The overlap in the right panel is related to the size of the direction sets and should not be interpreted as independent confirmation from two separate metrics.

A Representation Quality and Metric Interpretation

A.1 SAE quality and layerwise conventions

Figure A1 reports the validation explained variance and fraction of inactive features for each layer checkpoint. SAE quality varied across layers, so changes in candidate counts cannot be attributed solely to the visual-computation hierarchy. Failure to detect a candidate under strict criteria means that it did not meet the criteria for the current dictionary and control set; it does not mean that the layer contains no related information. Causal analyses at L22 used the 50,000-step checkpoint. The current results did not include retraining across multiple random seeds.

A.2 Full-set and Top-3 summaries

Figure A2 compares Duality computed from the full-set mean and Top-3 features. For larger cue sets, the two summaries can reflect different degrees of evidence concentration; direction sets contained at most two members, so Top-3 and full-set summaries used the same members. Sets with no members remained missing, and lines were not drawn across missing layers. The Duality reference line is a visual aid, not a significance threshold. Bias is consistently defined as convex evidence minus concave evidence.

Table B1: Conditions in the full scan that meet all three criteria. The median and small-change count refer to qualifying records in that row, not to all 696 pairs. The small-change threshold is computed from four typical FP16 steps before rounding.

Operation	Dose	Qualifying conditions	Small changes	Median |U effect|
Ablation	0.250	5	5	0.008
Ablation	0.500	16	15	0.008
Ablation	0.750	17	11	0.023
Ablation	1.000	31	14	0.055
Steering	0.500	15	15	0.008
Steering	1.000	27	22	0.008
Steering	2.000	30	22	0.008
Steering	4.000	22	12	0.027
Steering	8.000	27	16	0.023
Steering	16.000	10	6	0.027

B Full Scan and Same-Direction Paths

B.1 Scan conditions and numerical scale

Zero-dose conditions, $\alpha = 0.000$ and $\beta = 0.000$, left the selected contribution unchanged and yielded zero change. They were not counted among the 6,960 nonzero scan conditions. Table 2 reports five nonzero interventions; its E^+ L07 steering and E^+ L10 ablation values were +0.0128 and -0.0262 before rounding, respectively. Neither is a null effect. This distinguishes zero intervention dose from a small observed change under a nonzero operation.

$\epsilon = 10^{-6}$ is the floor of $\max(10^{-6}, 5\Delta_{\text{null}})$ with $\Delta_{\text{null}} = 0.000$, chosen because it lies almost four orders of magnitude below one FP16 margin step ($1/128 \approx 0.008$) and above float32 accumulation error, so it can only register numerical identity, never alter a change or reversal decision.

The full scan contained 696 deduplicated earlier-cue/later-direction pairs. Each pair was evaluated at six steering strengths and four ablation strengths, giving 6,960 nonzero conditions. Of these, 200 simultaneously met three criteria: the earlier intervention changed the later candidate, changed the output, and had a directionally consistent cross-layer contribution. Because records shared stimuli, features, and doses, they were not independent observations. Table B1 reports qualifying counts by operation and strength, together with counts of small changes.

Among the 200 qualifying conditions, 138 had an absolute earlier-intervention output change no larger than four typical FP16 margin steps, displayed as 0.031. Counts used the unrounded threshold. Directional consistency means that a condition passed the original numerical tolerance; it does not by itself imply a large effect. Each path should be interpreted alongside its absolute change and answer-mapping results.

B.2 Complete same-direction path values

At $\alpha = 4.000$, 15 same-direction paths met all three criteria: 12 concave-directed and 3 convex-directed. Table B2 reports all three intervention outcomes for every path, with zero change as the baseline. The largest concave transfer change was 0.133 for C^- L11 $\rightarrow$ E^- L12; the largest convex U-steering change was 0.023 for T^+ L12 $\rightarrow$ E^+ L15. These properties motivated the retrospective examples in Figure 5. They were not confirmatory tests independent of the scan.

B.3 Reversed ablation effects and local contributions

For the L14 cross-cue convex Top-3 set, target-directed changes at $\beta = 0.500$, 0.750, and 1.000 were +0.016, +0.023, and +0.031, respectively, indicating a stronger convexity bias after ablation. The Top-1 effect at $\beta = 1.000$ was +0.008. Decomposition showed that ablating the newly added member f44926 alone produced an approximately +0.039 convexity-directed change. These conditions did

Table B2: The 15 same-direction paths at $\alpha = 4.000$. All Δmargin values are in the direction of the path label; the baseline is 0.000. None of the paths shows directional consistency across both option mappings. Displayed 0.000 values include signed zeros in the source, which are rendered without a negative sign.

Path	Selection	Steer U	Steer U + restore R	Transfer R only
C^- L03 $\rightarrow$ E^- L11	Top-3	+0.008	0.000	+0.008
C^- L06 $\rightarrow$ E^- L10	Top-1	+0.016	+0.008	+0.008
C^- L07 $\rightarrow$ E^- L12	Top-1	+0.109	+0.102	+0.008
C^- L09 $\rightarrow$ E^- L11	Top-3	+0.031	+0.023	+0.016
C^- L09 $\rightarrow$ E^- L13	Top-1	+0.008	0.000	+0.008
C^- L10 $\rightarrow$ E^- L11	Top-1	+0.102	-0.039	+0.086
C^- L10 $\rightarrow$ E^- L12	Top-1	+0.102	-0.023	+0.070
C^- L10 $\rightarrow$ E^- L16	Top-1	+0.102	+0.094	+0.016
C^- L10 $\rightarrow$ E^- L20	Top-1	+0.102	+0.094	+0.008
C^- L11 $\rightarrow$ E^- L12	Top-3	+0.172	+0.039	+0.133
C^- L11 $\rightarrow$ E^- L16	Top-3	+0.172	+0.164	+0.016
C^- L12 $\rightarrow$ E^- L20	Top-3	+0.188	+0.180	+0.008
T^+ L05 $\rightarrow$ E^+ L07	Top-1	+0.008	0.000	+0.008
T^+ L05 $\rightarrow$ E^+ L10	Top-1	+0.008	0.000	+0.008
T^+ L12 $\rightarrow$ E^+ L15	Top-3	+0.023	+0.008	+0.008

not reverse the sign of the margin and did not produce a consistent reversal across both option mappings.

Spatial and activation records showed that only one patch of f65427 was active on the target image. Although f44926 was active on more patches, its image-level standardized activity remained below the convex-control mean. Thus, convex selectivity on the control set did not determine the local causal sign of the feature in every target state. SAE interventions pass through subsequent nonlinear computations; local contribution, feature selectivity, and final output direction must therefore be recorded separately. Work on activation replacement likewise showed that intervention effects should be interpreted with respect to the specific subspace and baseline (Makelov et al., 2024).

C Answer Mappings and Numerical Interpretation

For mapping 1, A denoted concave and B convex; mapping 2 reversed these assignments. The semantic margins were $m_1 = \ell$A,1 - ℓB,1 and $m_2 = \ell$B,2 - ℓA,2, with $M = (m_1 + m_2)/2$. A second baseline wording contrasted roof ridge (convex) with wall corner (concave). On that wording its ridge-minus-corner average was -0.281; the opposite sign would represent concave-minus-convex. In all four recorded baseline conditions the selected answer was the A-positioned option, so a positive semantic average was not a mapping-stable categorical preference. Mapping-specific baseline and intervention logits are needed to evaluate reversals. None of the 15 paths in Table B2 showed directional agreement across mappings.

All numerical changes in the main text and redrawn Figure 5 are reported to three decimals. The two updated E^+ entries in Table 2 are additionally given at source precision in the caption to prevent them from being mistaken for zero. Computation and thresholding used unrounded values. FP16 margins change in discrete steps, so small effects should be evaluated with raw logits, no-op conditions, and repeated-forward-pass tolerances. A distribution from matched random feature sets describes variation due to set sampling; it is not a population confidence interval across images or models.

D Cognitive and Computational Context

Separating input cues from depth representations clarifies how this study relates to broader depth research. Gibson emphasizes environmental structure available to an observer; Marr distinguishes the computational task from internal representations; and Palmer organizes problems in visual orga-

nization and depth perception (Gibson, 1979; Marr, 1982; Palmer, 1999). The effectiveness of cues also depends on viewing distance and context, so an effect of one cue should not be generalized to all depth tasks (Cutting & Vishton, 1995). Research on binocular distance perception addresses the formation of metric depth (Foley, 1980), whereas structure from motion and multiview geometry examine how constraints across views support 3D reconstruction (Ullman, 1979; Hartley & Zisserman, 2004). The present study examines concave–convex organization in a single image, a controlled component within this broader spectrum of abilities.

Human cue-combination studies provide a rationale for measuring the contribution of each cue separately. Weak-fusion accounts describe how cues combine as their scales and reliabilities change (Landy et al., 1995); visual–haptic experiments found integration consistent with statistical optimality under specific conditions (Ernst & Banks, 2002). Learning and adaptation also changed shape judgments based on illumination or other depth cues (Harding & Harris, 2012; Preston et al., 2009; Peveri et al., 2025). External frames affected stereoscopic depth magnitude, and conflicts between occlusion order and binocular disparity altered depth matching (Aida et al., 2024; Au et al., 2026). These findings motivate controlled cue comparisons. They do not imply that an SAE evidence score estimates the human weight assigned to that cue.

The transformation from cues to more stable depth representations also has computational analogues. Wundari et al. (2025) combined neuroimaging and a stereo-vision network to study the transition from local binocular correlations to more reliable matching representations; Xu et al. (2026) showed that viewing geometry inferred from optic flow can modulate motion and depth computations. Peveri et al. (2025) found that adaptation to cue-dependent invariants during active exploration changed depth perception. These studies involve different tasks and inputs and do not directly validate the model mechanism studied here, but they suggest concrete extensions involving viewpoint, cue reliability, and active sampling. The present conflict probe retains a basic experimental principle: manipulate evidence separately and compare its effects on internal representations and behavioral outputs.

E Broader Evaluation and Representation Literature

The broader literature supplies complementary evaluation levels. SEED-Bench and MMBench cover multimodal task performance (Li et al., 2024; Liu et al., 2024), while PhysBench focuses on physical-world understanding (Chow et al., 2025). Cognitive and neuropsychological evaluations examined uneven visual capabilities, serial processing, and linguistic context (Schulze Buschoff et al., 2025; Tangtartharakul & Storrs, 2026; Budny et al., 2025; Liu et al., 2025a). IllusionBench+, three-dimensional illusion depth estimation, and controlled semantic-ambiguity generation offer other diagnostic stimulus families (Zhang et al., 2025; Yao et al., 2025; Hu et al., 2026). These tasks define different demands on a model and should not be treated as interchangeable measures of a single spatial ability.

Representation work spans direct depth regression (Eigen et al., 2014), text-based three-dimensional learning (Cai et al., 2026), dense depth prediction (Yu et al., 2026), reconstruction-informed models (Fan et al., 2026), and abstract three-dimensional scene interfaces (Liu et al., 2026a). Probing studies examined geometric and scene information in visual foundation models (El Banani et al., 2024; Zhan et al., 2024; Man et al., 2024). Eyes Wide Shut, SSR, and depth-ordinal prompting considered visual limitations or alternative routes for supplying spatial information (Tong et al., 2024; Liu et al., 2025b; Vo et al., 2026). DepthVLA and post-training depth diagnostics extended these questions to action models (Yuan et al., 2025; Hackett et al., 2026). This body of work locates depth information at several computational interfaces; identifying which information affects one answer remains a separate intervention question.

Visual steering work provided precedents for manipulating SAE coordinates in CLIP (Joseph et al., 2025). RouteSAE addressed feature organization across layers (Shi et al., 2025), but shared dictionary organization alone does not establish the direction of causal influence. The present replacement test instead changed a specified later contribution after an earlier intervention and measured its effect on the same output. Its interpretation remains conditional on the selected dictionary, baseline, and answer mapping.