PRIOR DIRECTIONS: SEPARATING GEOMETRIC IDENTIFICATION FROM BEHAVIORAL EFFICACY IN VISUAL REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can remain anchored to stale verbal priors even when current visual evidence supports a different answer. We use this controlled revision failure to ask whether the subspace that best captures recurrent prior-induced representation change is also uniquely more behaviorally effective. Across four models, paired prior and reference states define Prior Directions, a compact low-rank geometry that, at rank 32, spans only 0.78% of the state dimensions yet captures 49.7% to 64.0% of held-out displacement energy. Matched correspondence improves held-out capture at every tested rank in all four models; across 20 permutation seeds, this advantage remains positive in all 15 available model–rank settings. Yet the geometric advantage does not translate into a comparable intervention advantage: matched and Pairing-Permuted edits restore $60/66$ and $59/66$ clean prior-induced failures, respectively. We trace this discrepancy to the geometry of the realized edits. Pairing-Permuted edits remain strongly aligned with the matched Prior Directions; their aligned component restores $59/66$ cases, whereas an equally norm-matched orthogonal residual restores only $20/66$. These results show that the construction that best identifies recurrent representation change need not be uniquely more behaviorally effective, distinguishing global subspace identification from the geometry actually used by an intervention.

1 INTRODUCTION

Vision-language models (VLMs) combine current visual input with verbal context from instructions, prior observations, or memory, as in interactive agents (Zeng et al., 2026). When the visual state changes, this context can become stale and redirect a correct judgment toward an outdated location. We call this directed revision failure *visual lock-in*.

Benchmarks document language-prior reliance and cross-modal conflict in VLM judgments (Lee et al., 2025, Deng et al., 2025, Golovanevsky et al., 2025a). We isolate this effect in a controlled grounding task derived from ScreenSpot-Pro (Li et al., 2025). The screenshot, query, and four answers stay fixed while verbal context changes. A stale cue names an outdated location, while irrelevant text controls for added language. Across four models, 69% to 99% of semantic-cue losses from initially correct judgments select the prior-consistent location (Figure 2), making prior-directed errors traceable under fixed visual evidence.

Paired pre-decision states under stale-prior and Irrelevant-Cue conditions define model-specific *Prior Directions*, connecting our analysis to visual task vectors, modality-gap geometry, and activation steering (Hojel et al., 2024, Yan et al., 2026, Saini et al., 2026, Wu et al., 2025). The directions generalize to held-out examples. At rank 32, they span 0.78% of the 4096-dimensional state and capture 49.7% to 64.0% of held-out prior-induced displacement energy (Figure 3), compactly describing recurrent change across visual instances.

Prior Directions raise a mechanistic question. We call held-out recurrence *geometric identification* and intervention effects *behavioral efficacy*, two properties studied in activation-space steering and causal editing (Liu et al., 2025, Jiang et al., 2025, Yang et al., 2025b). They provide complementary evidence. A fitted subspace can capture unseen changes globally while an alternative yields simi-

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.

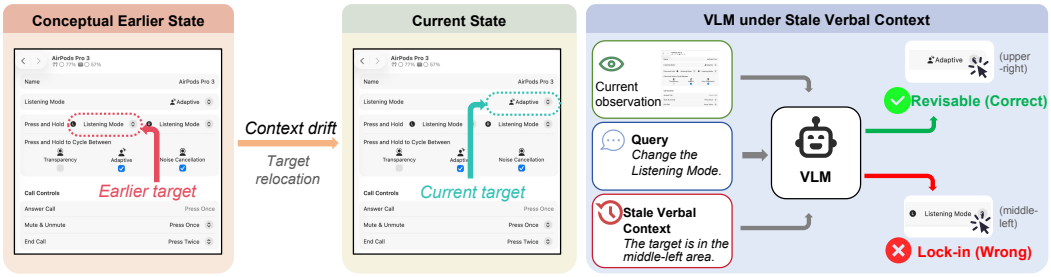

Figure 1. Schematic of our controlled visual-revision task with stale verbal context. The conceptual earlier state shows the location described by the verbal prior. Each model receives the current screenshot, query, answer choices, and verbal context. Revisable decisions select the current target, while visual lock-in selects the outdated location.

larly oriented edits on evaluated inputs. Global identification and input-conditioned edit geometry therefore require separate tests.

Following activation-space intervention studies (Wu et al., 2025, Liu et al., 2025), we compare matched and Pairing-Permuted constructions. Matched correspondence improves held-out capture at every tested rank in all four models, with a positive mean gap in all 15 settings across 20 permutation seeds (Figure 3). Intervention outcomes are nearly equal, restoring 60/66 and 59/66 clean visual-lock-in cases (Figure 4 and Table 2). Although the fitted subspaces overlap moderately, Pairing-Permuted edits remain aligned with matched Prior Directions on evaluated examples. Their aligned component restores 59/66 cases, compared with 20/66 for an equally norm-matched orthogonal residual. This locates most retained efficacy in the orientation shared by realized edits within each model.

Together, these results make three contributions.

- **Prior Directions.** We identify a compact, model-specific geometry of stale-prior-induced revision. At rank 32, it spans 0.78% of state dimensions and captures 49.7% to 64.0% of held-out displacement energy across four VLMs.
- **Separating geometric identification from behavioral efficacy.** Matched correspondence improves capture in all 20 four-model rank comparisons and yields positive mean gaps in all 15 repeated-permutation settings across three models. The two constructions restore $60/66$ and $59/66$ cases.
- **Explaining the separation through realized edit geometry.** The aligned Pairing-Permuted edit restores $59/66$ cases, compared with $20/66$ for an equally norm-matched orthogonal residual.

2 Related Work

Representation geometry and intervention. Global directions, hallucination subspaces, and selective steering alter vision-language behavior in activation space (Liu et al., 2025, Yang et al., 2025a, Wu et al., 2026, Yin et al., 2026, Shu et al., 2026). Visual task vectors and modality-gap analyses characterize compact and persistent representation geometry (Hojel et al., 2024, Yan et al., 2026). Recent approaches remove text-aligned components or steer prior-related internal states (Saini et al., 2026, Wu et al., 2025). Probing, corruption, and intervention studies also locate cross-modal interactions associated with hallucination and grounding (Golovanevsky et al., 2025b, Hua et al., 2025, Jiang et al., 2025, Yang et al., 2025b). These studies motivate two empirical objects that merit separate measurement. Held-out recurrence concerns the complete fitted subspace and its behavior on unseen state changes. An intervention concerns the vector generated from that subspace on a particular input. Our diagnostic evaluates each object while varying the prior-to-reference correspondence used in fitting. We then project the realized edit into matched and orthogonal components under equal edit norm. This comparison characterizes how a construction can improve global identification while an alternative construction preserves a behaviorally effective orientation.

Cross-modal conflict and stale context. Benchmarks document language priors, multimodal inconsistency, and knowledge conflict (Lee et al., 2025, Yan et al., 2025, Jia et al., 2025, Singla et al., 2026, Luo et al., 2025). Visual counterfacts and conflicting prompts show that language can redirect decisions despite relevant images (Golovanevsky et al., 2025a, Deng et al., 2025, Rudman et al., 2026, Ortu et al., 2026, Khayatan et al., 2026). Perception-knowledge interventions further characterize visual defaults and prior override (Lietzow et al., 2026, Cheng et al., 2026). Verbal memory can carry earlier observations into later agent decisions (Zeng et al., 2026). Our synthetic stale location creates a controlled conflict with fixed current visual evidence, a shared answer set, and paired states. Initially correct decisions and a designated outdated distractor make the behavioral consequence observable.

3 A Controlled Testbed for Prior-Driven Visual Revision

The experiments use a common controlled grounding protocol and model-specific representation spaces.

Models. We evaluate InternVL3.5-8B, MiniCPM-V-4.5, Qwen3-VL-8B, and Holo2-8B (Wang et al., 2025, Yu et al., 2025, Bai et al., 2025, H Company, 2025). All four models enter the primary behavioral, recurrence, and intervention analyses. The repeated-permutation reanalysis uses the saved geometry states from InternVL, Qwen, and Holo. They span distinct multimodal model families while remaining comparable in scale.

Table 1. Cohorts and evaluation roles. Selection for the wording scan is model-specific. Recurrence and intervention use the geometry cohort with distinct fitting protocols.

Analysis	Cohort	Evaluation rule
Directed failure	1,257 instances per model	Losses conditional on No-Cue correctness
Wording scan	Model-specific susceptible cases	Seven ordinal conditions, details in Appendix A
Recurrence	500 pairs per model	Five folds, each with 400 construction and 100 held-out pairs
Intervention	One $400/100$ split per model	66 clean failures across four evaluation sets

Tasks and Conditions. We derive a four-choice controlled conflict protocol from *ScreenSpot-Pro* (Li et al., 2025), using its screenshots, queries, and GUI targets as current evidence. The source benchmark contains $1{,}581$ tasks from 23 professional applications across five operating systems. We construct $1{,}257$ instances by partitioning each screenshot into a 3×3 grid and selecting four candidate regions comprising the current target, an incorrect region designated as its earlier location, and two distractors. This transformation preserves the screenshot and query while designating a synthetic earlier location and a matched discrete decision set. Text supplies the synthetic prior about the earlier location.

Each instance is evaluated under the cue set $\mathcal{C} = \{0, \mathrm{irr}, \mathrm{hyp}, \mathrm{fact}\}$. These are No-Cue, Irrelevant-Cue, Hypothesis-Cue, and Fact-Cue. Irrelevant-Cue controls for additional text, while the semantic cues state the same outdated location tentatively or as fact. The screenshot, query, and candidates remain fixed, yielding $5{,}028$ evaluations per model. Greedy decoding returns one of A through D within eight tokens. Thus cue wording is the only experimental input that varies within an instance.

Separating discovery from evaluation. The analyses use different eligibility rules because they answer different questions (Table 1). The initial behavior experiment starts from all constructed instances and measures losses conditional on No-Cue correctness. The wording scan selects previously susceptible cases to characterize responses to verbal certainty. Geometry is fit on a cohort sampled before that behavioral selection. Intervention outcomes are evaluated on clean failures within a disjoint evaluation split. Each denominator therefore corresponds to its evaluation target, comprising baseline-correct judgments, unseen paired changes, or reference-correct prior-induced failures.

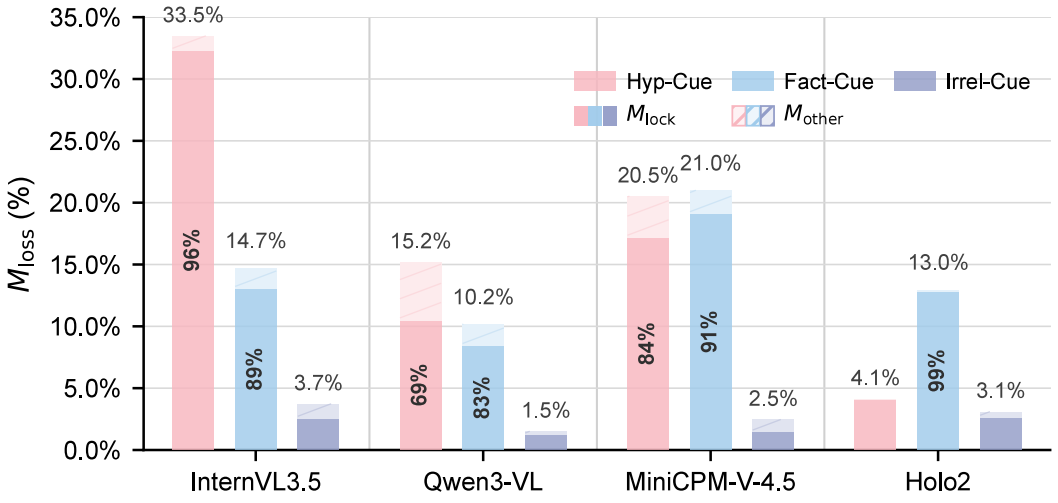

Figure 2. Prior cues redirect No-Cue-correct judgments. Bar height is the full-set lost mass M_{loss}. Solid M_{lock} and hatched M_{other} segments partition losses into prior-consistent and other errors. In-bar percentages give DCR, the prior-consistent share of losses. Gray bars show the Irrelevant-Cue control.

3.1 Directed Failures Validate the Testbed

We first test whether verbalized priors override correct visual judgments under fixed visual evidence.

Behavioral decomposition. A case is behaviorally eligible when its No-Cue judgment is correct. Let y_i be the visually supported answer, y_i^{lock} the prior-consistent distractor, and $\hat{y}_{m,i,c}$ the prediction of model m under cue c. The Lost-Correct Rate (LCR) and Directed Capture Rate (DCR) are

$$\begin{aligned} \text{LCR}_m(c) &= \Pr_i(\hat{y}_{m,i,c} \neq y_i \mid \hat{y}_{m,i,0} = y_i), \\ \text{DCR}_m(c) &= \Pr_i(\hat{y}_{m,i,c} = y_i^{\text{lock}} \mid \hat{y}_{m,i,0} = y_i, \hat{y}_{m,i,c} \neq y_i). \end{aligned} \tag{1}$$

LCR measures loss of a correct No-Cue judgment, and DCR measures the share directed to the prior-consistent option. The full-set masses are $M_{m,\text{loss}}(c) = A_{m,0}\text{LCR}_m(c)$ and $M_{m,\text{lock}}(c) = M_{m,\text{loss}}(c)\text{DCR}_m(c)$, where $A_{m,0}$ is No-Cue accuracy. Figure 2 partitions this loss mass by destination. The conditional revision analysis concerns initially correct judgments, and $A_{m,0}$ supplies their mass in the full test set.

Semantic priors induce directed lock-in. Figure 2 shows that semantic cues overturn substantially more No-Cue-correct judgments than the Irrelevant-Cue control. Semantic-cue loss ranges from 4.1% to 33.5%, while Irrelevant-Cue loss remains between 1.5% and 3.7%. Moreover, 69% to 99% of semantic-cue losses are directed to the prior-consistent option. The contrast identifies prior meaning as the organizing factor.

Standardizing the conflict condition. We additionally evaluate seven ordinal verbal-certainty conditions, $\rho \in \{-3, -2, -1, 0, +1, +2, +3\}$, all naming the same incorrect location. The index specifies an ordinal wording condition. We use its highest-certainty endpoint, $\rho = +3$, as the standardized condition for geometry. Empirically, this endpoint also yields the largest directed lock-in in each model, with model-specific intermediate trajectories. Representative prompts, cohort selection, and the scan appear in Appendix A. The scan uses behavior-selected model-specific cohorts. The recurrence analysis uses 500 pairs per model sampled before behavioral filtering.

4 Geometric Identification Through Recurrent Change

A recurring subspace would show that stale-prior influence has structure shared across examples, even when their screenshots and target regions differ. We measure this structure at the *pre-decision*

position, the final input token before answer generation. This position provides a common semantic anchor across answer tokenizations. Let $z^{(\ell)}_{m,i,+3}$ and $z^{(\ell)}_{m,i,\mathrm{irr}}$ be the hidden states for model m, example i, and layer ℓ under the prior and Irrelevant-Cue reference. Their paired difference is

$$\Delta z^{(\ell)}_{m,i} = z^{(\ell)}_{m,i,+3} - z^{(\ell)}_{m,i,\mathrm{irr}}. \tag{2}$$

The reference retains additional text without the prior-specific content. We use late layers $\mathcal{L}_{\mathrm{late}} = \{24, \ldots, 36\}$ throughout. Directions are fit separately for each model and layer. Recurrence refers to transfer across examples within the representation space of each model.

Constructing Prior Directions. For a construction set $\mathcal{T}$, we center its displacements and retain the top k right singular vectors.

$$\begin{gathered}\mu^{(\ell)}_{m,\mathcal{T}} = |\mathcal{T}|^{-1} \sum_{i \in \mathcal{T}} \Delta z^{(\ell)}_{m,i}, \qquad \widetilde{Z}^{(\ell)}_{m,\mathcal{T}} = [(\Delta z^{(\ell)}_{m,i} - \mu^{(\ell)}_{m,\mathcal{T}})^\top]_{i \in \mathcal{T}}, \\ \widetilde{Z} = U \Sigma V^\top, \qquad B_k = [v_1, \ldots, v_k], \qquad P_k = B_k B_k^\top.\end{gathered} \tag{3}$$

The columns of B_k are the Prior Directions. Centering removes the condition-wide mean shift from fitting, and SVD summarizes example-dependent variation. Held-out capture below projects the original, uncentered displacement, matching the intervention target. This distinction separates the fitting objective from the evaluation readout. Figure 3(a) illustrates the paired construction.

Held-out recurrence. We test recurrence independently of the behavior-selected cohorts by fitting this construction on a separate geometry cohort $\mathcal{G}_m$ of 500 controlled pairs per model at $\rho = +3$. In fold f, we learn $P^{(\ell,k)}_{m,f}$ from 400 pairs in $\mathcal{T}_{m,f} \subset \mathcal{G}_m$ and evaluate it on the 100 disjoint pairs in $\mathcal{H}_{m,f}$, for $k \in \{2, 4, 8, 16, 32\}$. Five-fold fitting keeps construction and evaluation disjoint, and the rank sweep tests recurrence across dimensionalities. On each held-out fold, C^{disp} measures the captured fraction of displacement energy, with $\epsilon = 10^{-12}$ stabilizing all denominators.

$$C^{\mathrm{disp}}_{m,f}(k) = \operatorname*{avg}_{\substack{\ell \in \mathcal{L}_{\mathrm{late}} \\ i \in \mathcal{H}_{m,f}}} \frac{\|P^{(\ell,k)}_{m,f} \Delta z^{(\ell)}_{m,i}\|_2^2}{\|\Delta z^{(\ell)}_{m,i}\|_2^2 + \epsilon}. \tag{4}$$

A rank-matched random subspace supplies an isotropic baseline. Pairing-Permuted fits the same construction after shuffling prior-to-reference correspondence within each construction fold. Evaluation retains the original matched held-out differences. Both controls keep rank, folds, and layers fixed.

At $k = 32$, which spans only 0.78% of the 4096-dimensional decision state, the matched Prior Directions capture 0.497 to 0.640 of held-out displacement energy, compared with 0.383 to 0.511 for Pairing-Permuted (Figure 3(b)). The matched subspaces exceed random controls across the rank sweep (Appendix B). A random rank-32 subspace has expected capture near $32/4096 = 0.0078$, far below the observed matched values. Recurrence is therefore substantial beyond dimensionality alone.

Recurrence across examples. The evaluation averages a normalized energy fraction for each example and layer, giving each observation equal weight. The fitted basis summarizes centered variation among construction pairs. Held-out projection measures how much of a fresh, uncentered change falls in that basis. These complementary choices make recurrence an empirical property of unseen changes. The captured fraction reports the geometry available to a subsequent edit, and the intervention tests its behavioral effect.

5 Matched Correspondence Improves Identification

The next test isolates the contribution of example correspondence to recurrence. Pairing-Permuted provides a targeted test by preserving the state collections and shuffling their correspondence. For centered construction-state matrices A, H and a permutation matrix Π,

$$\begin{aligned}X^\top X &= A^\top A + H^\top H - A^\top H - H^\top A, \\ X_\Pi^\top X_\Pi &= A^\top A + H^\top H - A^\top \Pi^\top H - H^\top \Pi A,\end{aligned} \tag{5}$$

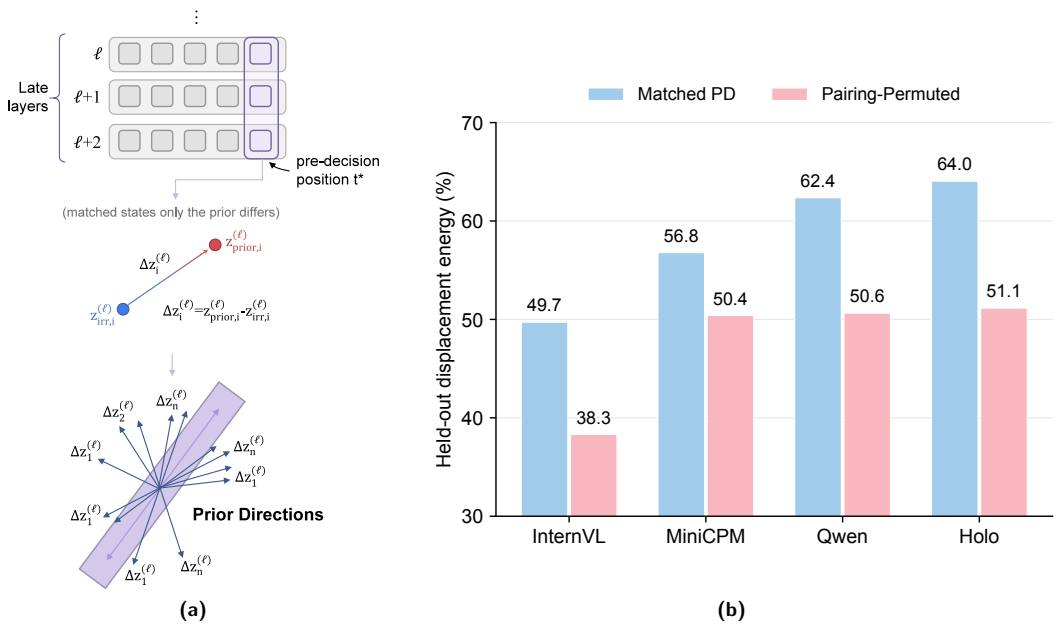

Figure 3. Prior Directions as compact recurrent structure in representation change. (a) Illustrative structural schematic showing how paired pre-decision states define shared low-rank directions across examples. (b) Held-out displacement energy captured by rank 32 matched and Pairing-Permuted directions. Bar labels show the absolute capture percentage for each model, and the isotropic expectation $k/d = 0.78\%$ lies below the displayed range. Matched correspondence exceeds Pairing-Permuted construction in all 20 four-model rank settings across the sweep. Across 20 permutation seeds, the mean gap is positive in all 15 tested settings for three models, as detailed in Appendix B.1.

where $X = A - H$ and $X_\Pi = \Pi A - H$. The marginal terms are unchanged. The cross-covariance terms change. Under a uniform permutation their expectation is zero after centering (Appendix D). This identity isolates correspondence as the construction-level difference between the two fitted subspaces.

A consistent gain across ranks and permutation seeds. In the original four-model sweep, matched capture exceeds Pairing-Permuted in all 20 model and rank comparisons (Appendix B). We repeat the permutation analysis with 20 seeds for the three models with saved per-example states, comprising InternVL, Qwen, and Holo. Each model uses the same 500 pairs, five folds, split seed 42, late layers, and ranks as above. All 15 settings retain a positive matched-minus-permutation-mean gap. Every sampled permutation lies below its matched value.

At rank 32, the gain is 0.114, 0.117, and 0.129 for InternVL, Qwen, and Holo, respectively. Across settings, permutation SD ranges from 0.00039 to 0.00257. Even the smallest mean gap, Qwen at rank 2, is positive at 0.008. The SD measures sensitivity to correspondence shuffling on the fixed cohort. The repeated-seed analysis covers the three models with saved per-example geometry states. Full values and numerical reproduction checks appear in Appendix B.1.

Correspondence contributes at the intervention rank. At rank 16, the repeated-seed gaps are 0.101, 0.105, and 0.124 for InternVL, Qwen, and Holo. The geometric advantage therefore also holds at the dimensionality used for intervention. Each permuted subspace is refit before evaluation, so this comparison includes the effect of the shuffled covariance on the learned directions. The resulting advantage identifies recurrent structure supplied by matched relations beyond the marginal geometry retained by the control.

Matched correspondence yields consistently higher held-out capture, while the permuted construction retains substantial capture. Behavioral efficacy requires a separate comparison of the edits derived from these constructions on the clean failure cohort.

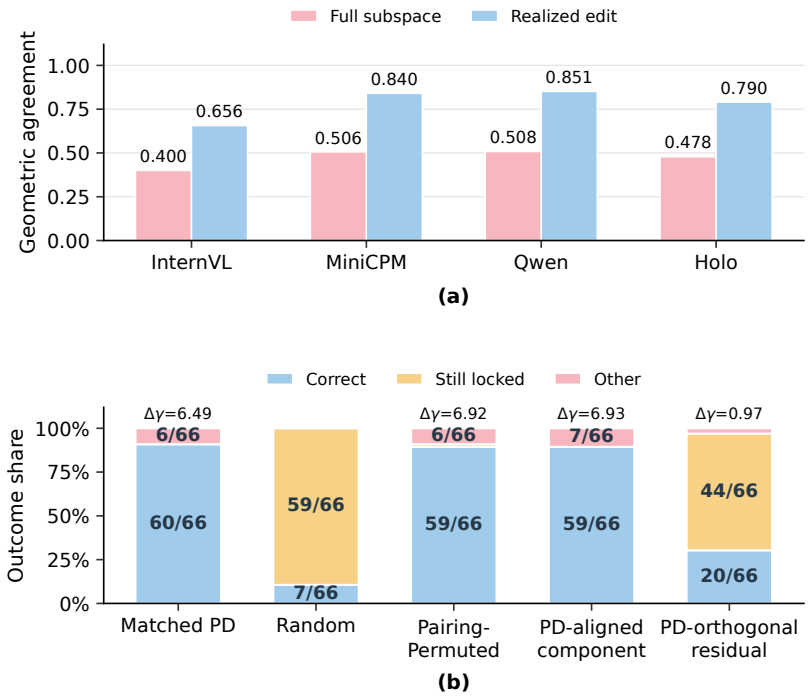

Figure 4. Behavioral efficacy is preserved along shared edit orientation within each model. (a) Global projector overlap S_{sub} and realized-edit alignment S_{eff}, both relative to matched Prior Directions. (b) Outcomes on 66 clean cases at $k = 16$ and $\lambda = 0.25$. $\Delta\gamma$ is the mean correct-versus-locked margin change. Random removal is norm-matched to the matched PD edit. Pairing-Permuted and its norm-matched aligned component retain most efficacy, while the norm-matched orthogonal residual has a weaker effect. Both components are defined relative to matched Prior Directions.

6 Behavioral Efficacy Is Preserved Along Shared Edit Orientation

Matched correspondence improves held-out capture, yet the two constructions produce almost the same restoration count. At the evaluated intervention strength, matched removal restores $60/66$ clean lock-in cases and Pairing-Permuted restores $59/66$. This separation between identification and efficacy motivates a comparison of the edits each construction actually applies.

For intervention, each 500-pair geometry cohort is split once into 400 construction and 100 disjoint evaluation pairs. We fit a rank-16 subspace on the construction split. Across the four evaluation sets, 66 clean lock-in cases are correct under Irrelevant-Cue and redirected to y_i^{lock} at $\rho = +3$. This criterion isolates prior-induced failures relative to the same-example reference. The intervention denominator is conditional on Irrelevant-Cue correctness, consistent with the reference used to define the displacement.

For each held-out instance, we subtract the component of prior-induced displacement aligned with the Prior Directions.

$$z_{m,i}^{(\ell),\text{int}} = z_{m,i,+3}^{(\ell)} - \lambda P_m^{(\ell,k)} \Delta z_{m,i}^{(\ell)}, \qquad \lambda = 0.25. \tag{6}$$

Here $\Delta z_{m,i}^{(\ell)}$ uses the matched Irrelevant-Cue reference. Hooks edit only the pre-decision token in $\mathcal{L}_{\text{late}}$, preserving other positions and reference components. All models and controls share λ, rank, layers, and the clean-case rule. Random and decomposed controls match edit norms per instance and layer to isolate orientation. The subtraction uses reference activations from the same example

as a counterfactual diagnostic. All directional comparisons reported here use the common strength $\lambda = 0.25$.

For each edit, we measure answer restoration and the change in correct-versus-locked logit margin, $\Delta\gamma_{m,i} = \gamma_{m,i}^{\text{int}} - \gamma_{m,i}^{\text{orig}}$, where $\gamma_{m,i} = s_{m,i}(y_i) - s_{m,i}(y_i^{\text{lock}})$ and $s_{m,i}(y)$ is the option logit.

Matched removal leaves no cases locked, while Pairing-Permuted leaves one locked. Random removal with matched edit norm restores $7/66$ and leaves $59/66$ locked, providing a directional control at equal edit magnitude. The similar matched and permuted restoration counts coexist with a clear recurrence gap. To locate the retained effect, we compare global projector agreement S_{sub} for the matched projector P and Pairing-Permuted projector P^{perm} with the fraction of realized Pairing-Permuted edit energy aligned with the Prior Directions, S_{eff}.

$$S_{\text{sub}} = \frac{1}{k}\,\text{tr}(PP^{\text{perm}})\,, \qquad S_{\text{eff}} = \frac{\|PP^{\text{perm}}\Delta z\|_2^2}{\|P^{\text{perm}}\Delta z\|_2^2 + \epsilon}. \tag{7}$$

Across models, S_{sub} ranges from 0.400 to 0.508, with median principal angles of $42.7°$ to $53.5°$, while S_{eff} reaches 0.656 to 0.851 (Figure 4(a)). Thus the realized edit retains substantial alignment despite differences between the full subspaces.

Global geometry and realized edits. Global subspace overlap gives every direction equal weight. Realized-edit alignment weights directions by the energy actually present in the edit applied to an example. A Pairing-Permuted subspace can therefore differ from the matched subspace globally while producing an edit that lies largely along matched directions. The principal-angle identity in Appendix D formalizes this weighting. The observed alignment calls for a directional intervention that separates the shared component from its orthogonal residual.

We decompose this edit into components aligned with and orthogonal to matched Prior Directions, and match each component to the norm of $P\Delta z$.

$$\begin{aligned} u_{\text{perm}} &= P^{\text{perm}}\Delta z, & u_{\parallel} &= \mathcal{N}(Pu_{\text{perm}}), \\ u_{\perp} &= \mathcal{N}((I-P)u_{\text{perm}}), & \mathcal{N}(v) &= \frac{\|P\Delta z\|_2}{\|v\|_2 + \epsilon} v. \end{aligned} \tag{8}$$

The aligned component restores $59/66$, leaves none locked, and raises the margin by 6.93 on average. The equally norm-matched orthogonal residual restores $20/66$, leaves $44/66$ locked, and raises the margin by only 0.97. This ordering holds in each model, with the full breakdown in Table 2. These norm-matched controls locate most retained efficacy in the component aligned with the matched Prior Directions. The orthogonal residual also restores some cases.

Consistent directional effects within each model. The aligned component restores $6/11$ InternVL cases, $21/23$ MiniCPM cases, $13/13$ Qwen cases, and $19/19$ Holo cases. The corresponding orthogonal counts are $3/11$, $7/23$, $2/13$, and $8/19$. Thus the aggregate restoration contrast is accompanied by the same directional ordering in every model. Matched and permuted removal also have closely spaced restoration counts within each model, differing by one InternVL case and yielding equal counts in the other three models. Each comparison uses a common clean cohort within its model, keeping eligibility fixed across the tested edits.

Table 2. Intervention outcomes by model. Each entry gives the number correct, still locked, and other, in that order. Aligned and orthogonal refer to components of the Pairing-Permuted edit relative to the matched Prior Directions within each model.

Model	Clean cases	Matched	Permuted	Aligned	Orthogonal
InternVL	11	7 / 0 / 4	6 / 1 / 4	6 / 0 / 5	3 / 8 / 0
MiniCPM	23	21 / 0 / 2	21 / 0 / 2	21 / 0 / 2	7 / 14 / 2
Qwen	13	13 / 0 / 0	13 / 0 / 0	13 / 0 / 0	2 / 11 / 0
Holo	19	19 / 0 / 0	19 / 0 / 0	19 / 0 / 0	8 / 11 / 0
Total	66	60 / 0 / 6	59 / 1 / 6	59 / 0 / 7	20 / 44 / 2

The outcome decomposition further separates releasing a prior-consistent answer from recovering the visual answer. Matched removal and the aligned component leave zero cases locked, with six and seven other errors, respectively. The orthogonal residual leaves 44 cases locked. In InternVL, the aligned component restores six cases and moves the remaining five to other answers. In Qwen and Holo, its restored counts cover the entire clean cohort. These model-specific outcomes localize the consistent effect to weakened prior-directed decisions, with the final recovery of the visual answer varying across the evaluated models.

Two geometries in a mechanistic diagnosis. Held-out displacement capture measures how well a fitted construction describes recurring change on unseen examples. Effective overlap and component removal measure the realized geometry of an edit on a specific input. Matched pairing improves the first property across the evaluated ranks and models. Pairing-Permuted edits retain much of the second property through directions aligned with the matched subspace, and this component preserves most of their restoration effect. Both levels of geometry are needed to interpret the intervention.

A diagnostic sequence for representation analysis. Each control isolates a different relationship between representation geometry and behavior. Rank-matched random subspaces measure capture attributable to dimensionality alone. Shuffling prior and reference correspondence tests what paired construction adds to recurrence while retaining the marginal state collections. Norm-matched random removal then tests the effect of edit magnitude, and the aligned and orthogonal components identify which part of the realized permuted edit carries the observed restoration. Together, these comparisons locate a correspondence-sensitive gain in global identification and an orientation-sensitive behavioral effect. The same sequence can guide interpretation of low-rank interventions in other controlled settings, where the relevant evidence must be measured anew.

7 Scope and Limitations

The evidence concerns synthetic stale verbal context under fixed current screenshots, four model families, and standardized English prompts. The single-image protocol isolates revision under controlled conflict. Extending this result to naturally accumulated memory and sequential interaction requires evaluations with observed histories. Recurrence and shared orientation are defined within each model, and broader task transfer remains an open empirical question.

Intervention efficacy is conditional on 66 clean failures, the specified reference, late layers, rank 16, and strength 0.25. The reference supplies same-example counterfactual information, making the edits suitable for mechanistic diagnosis. Their restoration counts characterize this evaluated cohort. Permutation SD measures randomness in construction on fixed examples, with ranks and folds sharing data. The experiments establish a correspondence advantage in recurrence and a dominant aligned component among the tested interventions.

8 Conclusion

Stale verbal priors provide a controlled setting in which representation changes are compact and recurrent. Matched correspondence improves the identification of those changes on unseen examples, while Pairing-Permuted edits retain similar restoration efficacy at the tested operating point. Their realized alignment and directional decomposition locate most retained efficacy in an orientation shared within each model. Prior Directions thus expose a useful diagnostic distinction. Geometric identification and behavioral efficacy are complementary tests of a representation-level mechanism, and each requires its own evidence.

AI USE STATEMENT

Generative AI tools assisted with polishing prose, retrieving and discovering related literature, research ideation and execution, and developing proofs for mathematical claims. The authors reviewed the sources, verified analyses and code against the experiments and implementation, independently checked the mathematical reasoning, and take responsibility for all claims and materials.

REPRODUCIBILITY STATEMENT

Sections 3–6 and Appendices A–E specify the inputs, prompts, cohorts, splits, seeds, model settings, controls, and analyses. The source archive provides prompts, figure scripts, and numerical summaries, but no raw hidden states; rerunning inference requires ScreenSpot-Pro and model weights.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating, 2026. URL https://arxiv.org/abs/2605.24024.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, 2025. doi: 10.1109/CVPR52734.2025.00366.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

H Company. Holo2 - open foundation models for navigation and computer use agents, 2025. URL https://huggingface.co/collections/Hcompany/holo2.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIII*, pp. 257–273, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72774-0. doi: 10.1007/978-3-031-72775-7_15. URL https://doi.org/10.1007/978-3-031-72775-7_15.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities?, 2025. URL https://arxiv.org/abs/2507.01790.

Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, Xiaoye Qu, Dongrui Liu, Lizhen Cui, and Yuntao Du. Benchmarking multimodal knowledge conflict for large multimodal models, 2025. URL https://arxiv.org/abs/2505.19509.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9f14fb9acd243c13c95d4a490d1684ce-Paper-Conference.pdf.

Pegah Khayatan, Jayneel Parekh, Arnaud Dapogny, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. When prompts override vision: Prompt-induced hallucinations in lvlms, 2026. URL https://arxiv.org/abs/2604.21911.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.231. URL https://aclanthology.org/2025.findings-naacl.231/.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use, 2025. URL https://arxiv.org/abs/2504.07981.

Niclas Lietzow, Danielle Bitterman, Carsten Eickhoff, William Rudman, and Michal Golovanevsky. Vision-default, prior-override: Causal mechanisms of perception-knowledge conflict in vision-language models, 2026. URL https://arxiv.org/abs/2606.28273.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/luo25b.html.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.642. URL https://aclanthology.org/2026.acl-long.642/.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Carsten Eickhoff, Ritambhara Singh, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision–language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41894–41912, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1941. URL https://aclanthology.org/2026.acl-long.1941/.

Harshvardhan Saini, Samyak Jha, Yiming Tang, and Dianbo Liu. When language overwrites vision: Over-alignment and geometric debiasing in vision-language models, 2026. URL https://arxiv.org/abs/2605.08245.

Huizhen Shu, Xuying Li, Hongxu Lin, Wenjie Sun, and Hui Li. Steering vision-language models with joint sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.25657.

Pratham Singla, Shivank Garg, Vihan Singh, and Paras Chopra. Do vision-language models see or guess? measuring and reducing textual-prior reliance with a phrasing-controlled benchmark, 2026. URL https://arxiv.org/abs/2606.10400.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. Revis: Sparse latent steering to mitigate object hallucination in large vision-language models, 2026. URL https://arxiv.org/abs/2602.11824.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. SHARP: Steering hallucination in LVLMs via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.725. URL https://aclanthology.org/2025.emnlp-main.725/.

Hanqi Yan, Xiangxiang Cui, Lu Yin, Jindong Gu, Paul Pu Liang, Yulan He, and Yifei Wang. Beyond cross-modal alignment: Measuring and leveraging modality gap in vision-language models, 2026. URL https://arxiv.org/abs/2502.14888.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (MMIR): A new benchmark for multimodal reasoning models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18829–18845, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.964. URL https://aclanthology.org/2025.findings-acl.964/.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection, 2025a. URL https://arxiv.org/abs/2412.13817.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8001c3568152d134d821cd46d4d84768-Paper-Conference.pdf.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models, 2026. URL https://arxiv.org/abs/2602.21704.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Ziyun Zeng, Hang Hua, Bocheng Zou, Mu Cai, Rogerio Feris, and Jiebo Luo. Mementogui: Learning agentic multimodal memory control for long-horizon gui agents, 2026. URL https://arxiv.org/abs/2605.18652.

A Experimental Protocol and Verbal Context

Controlled inputs. Each instance fixes the screenshot, current instruction, four candidate regions, correct option, and designated outdated option. The model receives the current screenshot and a prompt specifying the verbal context, instruction, and candidate answers. It returns one letter from A through D using greedy decoding with at most eight generated tokens. The No-Cue control omits the previous-context block. Irrelevant-Cue supplies a generic sentence about unrelated interface controls, application settings, or layout preferences. The fixed experiment seed selects this sentence after excluding messages containing the target-region or prior-region name.

Wording-scan eligibility. A pair enters the scan for model m when its original No-Cue response is correct and at least one semantic cue redirects it to the designated outdated option. Let $\mathcal{I}_m$ contain the selected pairs with complete records and Ω_m those that remain correct when No-Cue is reevaluated under the common scan wrapper. The sizes $(|\mathcal{I}_m|, |\Omega_m|)$ are $(417, 397)$, $(263, 238)$, $(159, 139)$, and $(165, 148)$ for InternVL, MiniCPM, Qwen, and Holo. Each selected pair produces nine records comprising No-Cue, Irrelevant-Cue, and seven verbal-certainty conditions.

Fixed wording selection. Each condition has three semantically matched templates. Representative sentences appear in Table 3. The placeholder $\{\texttt{region}\}$ denotes the same incorrect region in every condition for an instance. The template index is fixed by the pair identifier and the ordinal condition before model evaluation,

$$j(i, \rho) = 1 + \left(\sum_{c \in \texttt{pair_id}_i} \mathrm{ord}(c) + \rho \right) \bmod 3.$$

Thus wording selection is independent of the model output. The index spans $\rho \in \{-3, -2, -1, 0, +1, +2, +3\}$, progressing from uncertain recollection to explicit certainty about the same outdated location.

Behavioral readout. The directed lock-in rate over the complete selected cohort is

$$M_{m,\mathrm{lock}}(\rho) = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \Omega_m} \mathbf{1}[\hat{y}_{m,i,\rho} = y_i^{\mathrm{lock}}].$$

This denominator expresses the baseline-correct fraction multiplied by its conditional redirection rate. Table 4 reports all seven conditions. Intermediate responses vary by model. The highest-certainty endpoint attains the largest directed lock-in in each model and supplies the standardized geometry condition.

Table 3. Representative verbal-certainty templates. All conditions name the same outdated region. The complete three-variant template set and original wrapper are preserved in the accompanying source artifact.

ρ	Representative prior sentence
-3	With low confidence, I would say the earlier view maybe put the target in the $\{\texttt{region}\}$ region.
-2	A tentative reading of the earlier screen put the target in the $\{\texttt{region}\}$ region.
-1	The earlier screen gave a weak suggestion that the target might have been in the $\{\texttt{region}\}$ region.
0	The earlier note recorded the target location as the $\{\texttt{region}\}$ region.
$+1$	The earlier interface indicated that the target should be in the $\{\texttt{region}\}$ region.
$+2$	The earlier interface clearly showed the target located in the $\{\texttt{region}\}$ region.
$+3$	The earlier screen established that the target was in the $\{\texttt{region}\}$ region.

B Recurrence Protocol and Correspondence Controls

The geometry cohort contains 500 pairs per model sampled before behavioral filtering. Five-fold cross-fitting uses split seed 42, with 400 construction and 100 held-out pairs in each

Table 4. Directed lock-in across the ordinal wording scan. Entries report $M_{\text{lock}}(\rho)$. Intermediate trajectories are model-specific, while every model reaches its maximum at $\rho = +3$.

ρ	InternVL	MiniCPM	Qwen	Holo
-3	0.307	0.293	0.233	0.309
-2	0.372	0.452	0.447	0.339
-1	0.362	0.422	0.314	0.333
0	0.324	0.338	0.346	0.364
$+1$	0.388	0.525	0.453	0.400
$+2$	0.477	0.570	0.560	0.618
$+3$	0.501	0.605	0.610	0.800

Table 5. Prior-induced displacement recurs along the Prior Directions across rank. C^{disp} is the fraction of unseen prior-induced displacement energy captured by the fitted subspace. The matched Prior Directions exceed Pairing-Permuted in all 20 model and rank comparisons.

	InternVL		MiniCPM		Qwen		Holo	
k	Matched	Perm.	Matched	Perm.	Matched	Perm.	Matched	Perm.
2	**0.107**	0.046	**0.185**	0.177	**0.200**	0.191	**0.199**	0.171
4	**0.189**	0.093	**0.342**	0.315	**0.358**	0.333	**0.327**	0.284
8	**0.277**	0.164	**0.439**	0.388	**0.465**	0.401	**0.437**	0.358
16	**0.371**	0.270	**0.503**	0.442	**0.551**	0.447	**0.540**	0.415
32	**0.497**	0.383	**0.568**	0.504	**0.624**	0.506	**0.640**	0.511

fold. Each model and layer has its own centered SVD and projector. Evaluation averages normalized capture over held-out examples and layers 24 to 36. Table 5 gives the original matched and single-permutation results at all five ranks. At rank 32, random-subspace capture is $0.0077, 0.0078, 0.0078, 0.0078$ for InternVL, MiniCPM, Qwen, and Holo.

B.1 Repeated Permutations

The repeated-permutation analysis uses the saved geometry states of InternVL, Qwen, and Holo. It keeps the original cohorts, folds, layers, and ranks fixed and refits each permuted construction for 20 seeds. Evaluation always projects the matched held-out differences. Table 6 reports the results from the full-precision summary. SD quantifies variability across permutation seeds on these fixed examples. Every sampled permutation lies below matched capture in each of the 15 settings.

Recomputed matched capture agrees with the original results to maximum absolute error 2.44×10^{-7}, 1.12×10^{-7}, and 1.22×10^{-7} for Holo, InternVL, and Qwen. These checks establish numerical continuity with the original computation. MiniCPM contributes its original single-permutation results in Table 5.

C Intervention Outcomes by Model

Intervention uses one seed-42 split with 400 construction pairs and 100 evaluation pairs per model. A clean failure has a correct Irrelevant-Cue answer and a prior-consistent incorrect answer at $\rho = +3$. The resulting counts are 11, 23, 13, and 19 for InternVL, MiniCPM, Qwen, and Holo. All directional diagnostics use rank 16, layers 24 to 36, and strength $\lambda = 0.25$. Reference activations come from the same instance. Random removal and both decomposed components match the matched-PD edit norm separately for every instance and layer.

Table 2 in the main paper separates correct, locked, and other outcomes for each model. The aligned component restores more cases than the orthogonal residual in every model. Summing the model counts reproduces the main-paper totals of 60, 59, 59, and 20 restored cases. The reported outcomes are conditional restoration counts under this common diagnostic protocol.

Table 6. Repeated-permutation capture on fixed geometry cohorts. Gap is matched capture minus the permutation mean. SD measures variation over 20 permutation seeds.

Model	Rank	Matched	Perm. mean	Perm. SD	Gap
InternVL	2	0.107	0.046	0.00092	0.061
InternVL	4	0.189	0.094	0.00168	0.095
InternVL	8	0.277	0.163	0.00257	0.113
InternVL	16	0.371	0.270	0.00157	0.101
InternVL	32	0.497	0.384	0.00152	0.114
Qwen	2	0.200	0.192	0.00101	0.008
Qwen	4	0.358	0.334	0.00051	0.025
Qwen	8	0.465	0.401	0.00055	0.064
Qwen	16	0.551	0.447	0.00053	0.105
Qwen	32	0.624	0.507	0.00081	0.117
Holo	2	0.199	0.171	0.00195	0.028
Holo	4	0.327	0.284	0.00039	0.043
Holo	8	0.437	0.358	0.00075	0.079
Holo	16	0.540	0.416	0.00120	0.124
Holo	32	0.640	0.511	0.00139	0.129

D Projection and Intervention Identities

Let $B = [b_1, \ldots, b_k]$ and $C = [c_1, \ldots, c_k]$ be orthonormal bases for the matched and Pairing-Permuted subspaces within a model and layer, with projectors $P = BB^\top$ and $Q = P^{\text{perm}} = CC^\top$. All readouts and edits depend on these projectors, giving invariance to basis rotations, sign changes, and ordering.

Centered construction objective. Let the rows of X be $(\Delta z_i - \overline{\Delta z})^\top$, with SVD $X = U\Sigma V^\top$ and descending singular values σ_j. Write $B_k = [v_1, \ldots, v_k]$ and $P_k = B_k B_k^\top$. For any rank-k orthogonal projector R,

$$\|XR\|_F^2 = \operatorname{tr}(RX^\top X) \le \sum_{j=1}^{k} \sigma_j^2 = \|XP_k\|_F^2,$$
$$\|X - XR\|_F^2 = \|X\|_F^2 - \|XR\|_F^2 \ge \|X - XP_k\|_F^2.$$

Thus the fitted Prior Directions minimize squared reconstruction error on centered construction displacements. The empirical recurrence test evaluates the captured fraction of uncentered displacements on disjoint examples.

Permutation control. For centered prior and reference matrices $A, H \in \mathbb{R}^{n\times d}$, matched and permuted differences are $X = A - H$ and $X_\Pi = \Pi A - H$. A uniform permutation preserves both marginal covariance terms. Centering gives $\mathbb{E}[\Pi] = n^{-1}\mathbf{1}\mathbf{1}^\top$ and

$$\mathbb{E}_\Pi[X_\Pi^\top X_\Pi] = A^\top A + H^\top H.$$

This expectation specifies the covariance effect of disrupting correspondence. The reported capture is computed by fitting and evaluating each sampled permutation separately, retaining the nonlinear dependence of the leading eigenspace on the sampled covariance.

Global and realized overlap. Let θ_j denote principal angles between the two subspaces. Express the realized permuted edit as $u = Q\Delta z = \widetilde{C}\alpha$ in the principal basis of Q. For a nonzero edit,

$$S_{\text{sub}} = \frac{1}{k}\sum_{j=1}^{k} \cos^2\theta_j, \qquad \frac{\|Pu\|_2^2}{\|u\|_2^2} = \sum_{j=1}^{k} w_j \cos^2\theta_j, \qquad w_j = \frac{\alpha_j^2}{\sum_r \alpha_r^2}.$$

Global overlap weights each direction equally. Realized overlap weights by edit energy. The stabilized statistic multiplies the latter expression by $\|u\|_2^2/(\|u\|_2^2 + \epsilon)$. Concentrating edit energy along closely aligned principal directions yields high effective overlap even for moderately overlapping full subspaces.

Norm matching. For $a = \|P\Delta z\|_2$, the control uses $\mathcal{N}_\epsilon(v) = av/(\|v\|_2 + \epsilon)$, with $v = Pu$ or $v = (I-P)u$. This map sends zero to zero. For nonzero v, its norm has relative shortfall $\epsilon/(\|v\|_2+\epsilon)$ from a. Equal-strength edits therefore control magnitude up to the numerical stabilizer, leaving orientation as the designed difference. The downstream network evaluates these interventions at finite strength, and the restoration counts provide their behavioral readout.

E Computational Details

Original inference and state extraction used a single NVIDIA GeForce RTX 5090 GPU, CUDA, and `bfloat16` model states. The software stack comprised Python, PyTorch, Transformers, NumPy, and Pillow. Model-specific configurations determine hidden-state extraction and intervention hooks. Generation uses temperature 0, disabled sampling, and at most eight output tokens. Each model, condition, and instance combination is evaluated once.

The original configurations use seed 42. The repeated-permutation analysis additionally varies the correspondence-shuffling seed over 20 runs. It projects saved states using Apple MPS and performs small Gram eigendecompositions on CPU. Across analyses, the numerical stabilizer is $\epsilon = 10^{-12}$. Geometry uses ranks $2, 4, 8, 16, 32$, and the reported intervention comparison uses rank 16 and strength 0.25.