TESTING FINAL-STATE SUFFICIENCY IN MULTIMODAL GRAPH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning about changing worlds requires models to answer questions after a sequence of updates. A correct response, however, does not show that the model's prediction depends only on the reached state: different histories may arrive at the same endpoint yet change how strongly the model favors the correct answer over valid alternatives. Final-state sufficiency requires such histories to yield the same answer distribution for a fixed task. Consequently, the *correct-answer margin*—the correct complete answer's log-likelihood minus that of its strongest valid alternative—must also remain unchanged. We test this implication using matched graph histories that add the same two edges in reverse order while fixing the final graph, question, valid answers, and text or image presentation. Questions about unchanged graph properties provide matched controls for order sensitivity. In a controlled construction ending at the complete graph on four nodes, both Qwen3-VL and Gemma violate the required margin equality. A separate held-out study extends the Gemma result to four non-complete graph families with six or eight nodes, where independent calibration verifies task competence and adoption of both updates. These findings show why endpoint accuracy should be complemented by direct invariance tests across state-equivalent histories.

1 INTRODUCTION

Many reasoning tasks present a model with an initial state, provide a sequence of updates, and ask what is true afterward. Getting the final answer right does not show that the model's prediction was determined by the reached state alone. Two update histories can arrive at exactly the same endpoint while changing how strongly the model favors the correct answer over other valid answers. If the question depends only on that endpoint, the answer distribution should remain unchanged across such histories when the rest of the task is fixed. We call this requirement *final-state sufficiency*: a model's answer distribution may depend on a history only through the state it reaches.

Prior work on dynamic reasoning asks whether models can track changing entities, recover intermediate states, or encode latent world structure (Dalvi et al., 2018; Richardson et al., 2022; Li et al., 2023). Other studies show that changing the order of demonstrations, premises, or answer options can alter model predictions (Lu et al., 2022; Chen et al., 2024; Pezeshkpour & Hruschka, 2024). Evaluations of world-model coherence also test whether histories reaching the same state admit consistent future continuations (Vafa et al., 2024). Here we ask whether update order changes the answer evidence for a fixed question about the final state. Accuracy can miss such dependence because the correct answer may remain ranked first while its relative support changes. We therefore compare the correct complete answer with the valid alternative that the model favors most, calling their log-likelihood difference the *correct-answer margin*. The margin is positive exactly when the correct answer outranks every valid alternative; larger values indicate a greater log-likelihood advantage. Final-state sufficiency requires equal margins across histories reaching the same state. Unequal margins therefore refute the requirement, although equal margins do not establish identical answer distributions.

Graphs provide a controlled setting because their states and updates can be specified exactly in text or images (Wang et al., 2023; Li et al., 2024). Figure 1 shows our construction: starting from K_4, the complete graph on four nodes, with two incident edges removed, two histories add the same edges in opposite orders while fixing the final graph, question, valid answers, and presentation.

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.

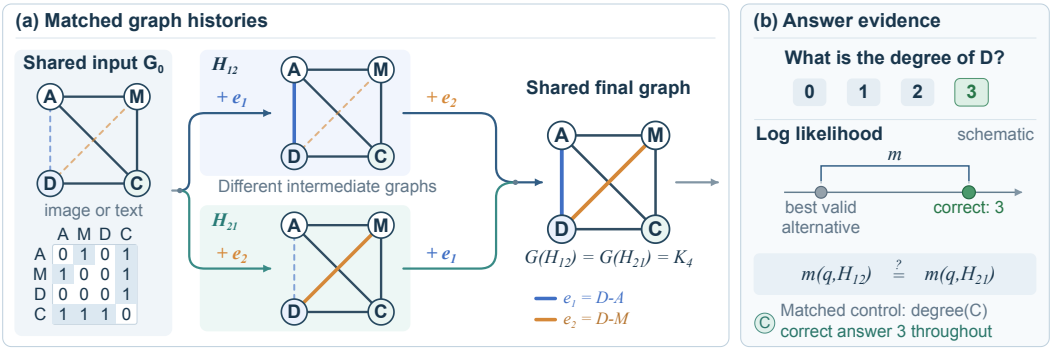

Figure 1: **Matched-history test of final-state sufficiency.** (a) One initial graph, shown as an image or adjacency matrix, receives the same two edge additions in opposite orders. Dashed colored segments mark absent edges; both histories reach K_4. (b) The degree question and valid answers are shared. The correct-answer margin m is the log-likelihood gap between the correct complete answer and its strongest valid alternative. Final-state sufficiency requires equal margins; the scores shown are schematic. The matched control has the unchanged answer $\deg(C) = 3$. The formal test combines control-adjusted contrasts across presentations and question families (Section 3.2).

We measure the margin difference between histories for a question about an updated property and subtract the corresponding difference for a matched question about an unchanged property. We use questions about node degree and shortest-path length. Our prespecified interaction asks whether the difference in adjusted order sensitivity between these question types changes across text and image presentations. A nonzero interaction implies that at least one underlying history comparison violates the required margin equality.

On held-out instances of the K_4 construction, the prespecified interaction is significantly above zero for both Qwen3-VL-8B-Instruct (Bai et al., 2025) and Gemma-4-12B-it (Gemma Team, 2026). Their answer evidence therefore depends on update order despite identical endpoints. To examine whether this finding depends on the dense K_4 endpoint, a separate held-out study tests five non-complete graph families at six and eight nodes. For Gemma, independent calibration confirms graph-question competence and adoption of both updates in seven family–size combinations. The joint equality of the control-adjusted order contrasts is rejected in all seven after multiplicity correction, covering four graph families and both tested sizes. The two studies provide complementary scope: the controlled result spans two models, while the extension shows that Gemma's history sensitivity is not specific to K_4.

To interpret this dependence, we test the effect of explicit final-state information and examine internal states. A separate diagnostic compares graph-changing and state-preserving instructions that reach the same endpoint. Adding a correct text or image description reduces the magnitude of their aggregate behavioral contrast in three of four combinations of model and description format. We then compare how hidden-state differences induced by update order vary across question conditions in the two models (*relational agreement*) (Kriegeskorte et al., 2008). Agreement is higher while processing the second update than immediately before the answer. To test whether these states affect the primary interaction, we replace selected activations with those from the opposite history (Geiger et al., 2021). Replacement has a larger effect at the pre-answer position, but the prespecified controls do not establish that this influence is selective for the relevant history–question relation.

Together, these studies distinguish endpoint correctness from invariance of answer evidence to the history that produced the endpoint. Matched state-equivalent histories make this distinction testable by holding the reached graph and answer task fixed while varying how the graph was reached. Evaluations of dynamic reasoning should therefore examine answer evidence across state-equivalent histories alongside endpoint accuracy.

2 Related Work

Recovering the current state is central to reasoning about a changing world. State-tracking work evaluates this ability in procedural text and controlled transition systems (Dalvi et al., 2018; Tandon et al., 2020; Kim & Schuster, 2023; Li et al., 2025). Related mechanistic studies recover latent world-state representations (Li et al., 2023; Feng et al., 2025), whereas behavioral studies test whether predictions remain consistent with the states and transitions of an environment (Vafa et al., 2024; Harang et al., 2025). Visual and multimodal graph benchmarks evaluate relational understanding and graph reasoning across rendered and textual graph representations (Li et al., 2024; Ai et al., 2024; Das et al., 2024; Babaiee et al., 2025; Sun et al., 2026). These studies motivate a distinction between recovering a state and relying on that state alone. We test the latter by holding the final graph fixed and comparing the correct-answer margin across state-equivalent histories.

Prompt-order studies reorder demonstrations, answer options, or context blocks (Lu et al., 2022; Pezeshkpour & Hruschka, 2024; Ok & Lee, 2026). Our histories instead pass through different intermediate graph states before reaching the same endpoint. Our diagnostic also differs from methods that remove order information to enforce permutation invariance (McIlroy-Young et al., 2024): we measure order dependence rather than eliminate it. To interpret history dependence, we draw on methods for comparing and intervening on internal states. Representational comparisons quantify relational agreement without explicit coordinate alignment (Kriegeskorte et al., 2008; Kornblith et al., 2019), whereas activation and interchange interventions test causal influence (Li et al., 2025; Meng et al., 2022; Geiger et al., 2024). We further ask whether that influence is specific to the relevant history–question relation. Appendix A expands on these connections.

3 Testing Final-State Sufficiency with Matched Graph Histories

We compare answer evidence across graph histories that reverse the same two updates while preserving the initial graph, final graph, and answer task.

3.1 Final-state sufficiency and the correct-answer margin

Consider a model given an initial graph G_0, a sequence of updates $h = (u_1, \dots, u_L)$, and a question about the resulting graph. Each u_j denotes the graph operation specified by an update, giving the final graph $G(h) = (u_L \circ \cdots \circ u_1)(G_0)$. We consider questions q whose correct answer depends only on this final graph. Final-state sufficiency requires the model's answer distribution to depend on h only through $G(h)$. Let $P(a \mid q, h)$ be the model's likelihood of a complete answer a, with the shared initial graph, presentation, and answer format left implicit. For two histories under this fixed task, the requirement is

$$G(h) = G(h') \quad \Longrightarrow \quad P(a \mid q, h) = P(a \mid q, h') \quad \text{for every } a \in \mathcal{A}(q), \tag{1}$$

where $\mathcal{A}(q)$ is the finite set of valid complete answers: the allowed responses, not just the correct one. The requirement concerns every candidate likelihood, not just the top-ranked answer.

Accuracy cannot detect a change in answer likelihoods when the correct answer remains ranked first. To measure its advantage over the closest competitor, we compare the correct complete answer $a^\star$ with the highest-likelihood valid alternative. Writing $\mathcal{A}^-(q) = \mathcal{A}(q) \setminus \{a^\star\}$, we define the *correct-answer margin* as

$$m(q, h) = \log P(a^\star \mid q, h) - \max_{a \in \mathcal{A}^-(q)} \log P(a \mid q, h). \tag{2}$$

The margin is positive exactly when the correct answer outranks every valid alternative, and larger values indicate a greater log-likelihood advantage. Thus Eq. (1) requires equal margins for histories reaching the same graph. Unequal margins refute final-state sufficiency, although equal margins do not establish equality of the full answer distributions.

3.2 Matched histories and the primary behavioral estimand

To vary update order without changing the initial or final graph, we begin with K_4, the complete graph on four nodes, and remove two edges e_1, e_2 that share an endpoint (Figure 1). History H_{12}

adds e_1 and then e_2; H_{21} adds them in reverse order. Each addition changes the graph when applied, and both histories end at K_4. Within a history pair, the initial graph presentation, question, and valid answers are fixed; only the two update statements exchange positions.

The initial graph is presented as a textual adjacency matrix or an image; updates and questions remain textual in both cases. Degree questions ask about local adjacency, whereas shortest-path questions ask about paths between nodes. Within each family, an *update-relevant* question concerns a property changed by the additions, and its *matched control* concerns a property that neither addition changes. In Figure 1, for example, the degree of D increases from one to three, while the degree of C remains three. The control lets us compare order sensitivity for a property that required updating with sensitivity for one that did not.

A *matched graph instance* groups both histories and all question and presentation conditions for one realization of node labels, questions, and layout. All behavioral quantities are computed separately for each model. Let $i \in \{1, \dots, 24\}$ index the held-out instances, $M \in \{T, I\}$ the initial text or image presentation, and $Q \in \{d, \mathrm{sp}\}$ the degree or shortest-path family. With $r \in \{A, U\}$ denoting the update-relevant or control role, write $m^r_{i,M,Q}(H)$ for the corresponding margin. We define the *order contrast* and its matched-control adjustment as

$$\begin{aligned} \Delta^r_{i,M,Q} &= m^r_{i,M,Q}(H_{21}) - m^r_{i,M,Q}(H_{12}), \\ E_{i,M,Q} &= \Delta^A_{i,M,Q} - \Delta^U_{i,M,Q}. \end{aligned} \tag{3}$$

Thus $\Delta^r_{i,M,Q}$ measures the signed change in margin when the updates are reversed, and $E_{i,M,Q}$ compares that change between the updated and unchanged properties. The instance construction and question-matching rules are given in Appendices B.1 and B.4.

The primary comparison asks whether presenting the graph as text or an image changes the difference in adjusted order sensitivity between degree and shortest-path questions. Within each presentation, we compare the two question families; we then compare these differences across presentations. This gives the prespecified interaction for each instance and its held-out mean:

$$\begin{aligned} \Psi_i &= \big(E_{i,T,d} - E_{i,T,\mathrm{sp}}\big) - \big(E_{i,I,d} - E_{i,I,\mathrm{sp}}\big), \\ \widehat{\theta} &= \frac{1}{24} \sum_{i=1}^{24} \Psi_i. \end{aligned} \tag{4}$$

A positive $\widehat{\theta}$ means that the degree-minus-shortest-path difference is greater under text than under images. Additive presentation and question-family effects cancel in this comparison.

Final-state sufficiency requires every $\Delta^r_{i,M,Q}$ to be zero, which in turn requires $\widehat{\theta} = 0$. A nonzero interaction therefore implies a violation in at least one underlying history comparison. Our directional hypothesis is **H1**: $\widehat{\theta} > 0$. Its positive orientation was selected on development examples and fixed before held-out evaluation. The relationship between this fixed linear contrast and the full set of margin equalities, together with the inference procedure, is detailed in Appendices D.1 and D.2.

3.3 Extending the equality test beyond K_4

The K_4 construction isolates update order at a fixed complete-graph endpoint. To examine whether history dependence also occurs at non-complete endpoints, we extend the paired construction to other graph families and sizes. The added edges play different structural roles, such as connecting components, closing a cycle, or providing an alternative path. Within each pair, the same two additions are reversed and the final graph and answer task remain fixed. We retain both presentations, both question families, and their matched controls.

Before interpreting history differences, we check whether a model can answer the graph questions and incorporate both requested updates. These checks use independent calibration examples, without selecting on the history contrast. For each eligible model–family–size combination, a joint test asks whether the four control-adjusted contrasts—text and image crossed with degree and shortest path—are all zero. This tests the same necessary invariance through a joint comparison rather than H1's directional interaction, allowing the four contrasts to have different signs. The graph constructions, calibration criteria, and joint inference are specified in Appendices F.1 to F.3.

3.4 FINAL-STATE DESCRIPTIONS AS A SEPARATE DIAGNOSTIC

We next ask whether explicitly stating the reached graph can reduce behavioral differences between histories that reach it. For this diagnostic, we compare two one-update conditions. The *graph-changing* condition adds an edge $e^\star$ to $K_4 \setminus \{e^\star\}$; the *state-preserving* condition requests the same addition when that edge is already present in K_4. Both reach K_4 and share the answer task, but only the first requires a change to the initial graph. The two conditions therefore differ in their initial graphs, not in update order as in H1.

The baseline provides no additional description (c_0). In the correct-description conditions, we append the final graph in text (c_T) or as an image (c_I) after the update and before the question. A further condition appends a stale image of $K_4 \setminus \{e^\star\}$ (c_S). Whenever a description is present, the prompt instructs the model to use it as the current graph. The correct descriptions test whether explicit endpoint information reduces the behavioral contrast. Comparing stale and correct images asks whether the remaining contrast is sensitive to the displayed image condition. Description format varies independently of whether the initial graph is presented as text or an image (Appendix B.3).

For each description condition c, we subtract the graph-changing margin from the state-preserving margin for an update-relevant question, then subtract the corresponding contrast for its matched control. We average these adjusted contrasts within each question family, then apply the presentation–question-family interaction in Eq. (4). Its mean over the 24 matched instances is denoted $\widehat{\theta}_c^{\mathrm{final}}$. This retains the primary study's comparison across presentations and question families, while replacing its order contrast with the contrast between the two one-update conditions. The full definition is given in Appendix D.1.

Attenuation concerns how close this aggregate is to zero, regardless of its sign. For a correct description in format $F \in \{T, I\}$, we measure the reduction from the no-description baseline and, separately, compare stale and correct images:

$$\begin{aligned} \widehat{\tau}_F &= \left|\widehat{\theta}_{c_0}^{\mathrm{final}}\right| - \left|\widehat{\theta}_{c_F}^{\mathrm{final}}\right|, \\ \widehat{\xi} &= \left|\widehat{\theta}_{c_S}^{\mathrm{final}}\right| - \left|\widehat{\theta}_{c_I}^{\mathrm{final}}\right|. \end{aligned} \tag{5}$$

H2, $\widehat{\tau}_F > 0$, tests whether a correct text or image description reduces the aggregate magnitude for each model. **H3**, $\widehat{\xi} > 0$, tests whether the stale image leaves a larger magnitude than the correct image. Absolute values are taken after averaging signed instance contrasts, so attenuation need not occur in every instance. Inference and its assumptions are specified in Appendix D.2.

3.5 INTERNAL EVIDENCE: AGREEMENT, INFLUENCE, AND SELECTIVITY

The internal analyses return to the core H_{12}/H_{21} histories. We ask whether the two models organize their internal history differences similarly, whether the selected states affect the primary behavioral contrast, and whether that influence is specific to the relevant history–question relation. These are distinct questions: agreement need not imply influence, and influence need not be selective. At a fixed decoder layer within each model, we examine residual-stream states at two locations: the tokens expressing the second update ($\ell = \mathrm{upd}$) and the final prompt token before the answer ($\ell = \mathrm{ans}$). This lets us examine both update processing and answer preparation; exact sites are given in Appendix C.2.

Relational agreement. We compare the relations among history differences across questions, rather than aligning individual coordinates between models (Kriegeskorte et al., 2008). The question is whether conditions with similar history differences in one model also have similar differences in the other. For each instance, presentation, and location, we average each model's selected token states and subtract the H_{12} vector from the H_{21} vector. Degree and shortest-path questions in the two roles yield four history-difference vectors per model.

Their six pairwise cosine dissimilarities form a representational dissimilarity matrix (RDM). Spearman correlation between corresponding entries measures cross-model relational agreement. For each presentation M, we average the paired second-update-minus-pre-answer correlation differences across eligible instances, obtaining $\widehat{\zeta}_M$. A positive value indicates greater agreement at the second-update span. The complete estimand and eligibility rules appear in Appendix D.4.

Causal influence. To test whether the selected states affect the primary behavioral contrast, we replace states from one history with those from its matched opposite history (Geiger et al., 2021). The target and donor share the model, graph instance, presentation, and question. We leave the target prompt unchanged and observe the resulting answer margins. Let H_s be the target and H_t the donor, with $(s, t) \in \{(12, 21), (21, 12)\}$. Writing $Z_\ell(H)$ for the full tensor of selected token states, we interpolate as

$$\widetilde{Z}_\ell^{s \leftarrow t}(\alpha) = (1 - \alpha) Z_\ell(H_s) + \alpha Z_\ell(H_t), \qquad \alpha \in [0, 1]. \tag{6}$$

Here $\alpha = 0$ retains the target state and $\alpha = 1$ replaces it with the donor state. Unlike the mean-pooled relational comparison, this operation acts on every selected token vector. Position and shape matching rules are specified in Appendix C.2.

For each question, we subtract the unmodified target margin from the intervened margin, reversing the sign for $21 \leftarrow 12$ to align both directions with $H_{21} - H_{12}$. We then apply the primary test's control adjustment and interaction and average over instances. This direction-aligned intervention contrast measures influence on the behavioral comparison. Dividing it by the observed $\widehat{\theta}$ gives a normalized scale on which one matches the original aggregate in size. The complete estimand and inference are given in Appendix D.5.

Selectivity controls. At the pre-answer position, we test whether full replacement depends on the matched history–question pairing rather than nonspecific disturbance. We measure movement on matched-control questions both in absolute terms and as a contribution to the signed interaction. We also compare matched replacement with norm-matched random perturbations, same-answer donors from another instance, and different-question donors from the same instance. These perturbations and alternative donors must have small effects relative to matched replacement. Identity reinsertion checks that the operation itself leaves the model unchanged.

A separate cross-family comparison requires positive transfer of an estimated history difference between question families, with limited disturbance on control questions. Selectivity requires every prespecified comparison to be available and to satisfy its criterion. Appendix D.6 defines the statistics, donor operations, and thresholds.

4 Experiments

We test whether update order changes answer evidence when the final graph and question are fixed, using the controlled K_4 construction and a held-out extension to larger graphs. We then examine how explicit final-state information changes the separate one-update contrast and compare internal relational agreement with intervention effects.

4.1 Evaluation setup

We evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) and Gemma-4-12B-it (Gemma Team, 2026), hereafter Qwen3-VL and Gemma, on 24 held-out matched graph instances in the core K_4 study. Each instance is evaluated under both update histories and both initial-graph presentations, with update-relevant and matched-control questions about degree and shortest-path length. We compute the correct-answer margin by scoring every valid complete answer rather than sampling responses. The final-state-description diagnostic uses 24 matched instances with the one-update design in Section 3.4. Prompt and model details appear in Appendices B and C.1.

The topology extension evaluates these models and Qwen3.5-9B (Qwen Team, 2026) on five non-complete graph families at node counts $n \in \{6, 8\}$. Each family–size template has 32 held-out variants of node labels and image layout. We keep all matched conditions from a variant together as one graph cluster for inference; the variants do not represent different topologies. Each model–family–size combination forms a panel. Independent calibration checks task competence, edge recognition, update adoption, and response parsing before a panel enters the history test. The constructions and calibration criteria are given in Appendices F.1 and F.2.

Internal analyses use the core K_4 histories at a fixed layer near 80% of decoder depth in each model. We compare the second-update span with the pre-answer position and apply selectivity controls at the latter (Appendix C.2). Throughout the core study, 95% bootstrap intervals resample matched

instances, keeping their paired conditions together. Confirmatory tests preserve this pairing and adjust for multiple comparisons within the prespecified families. The corresponding procedures for the topology extension operate on its graph clusters (Appendices D.2 and F.3).

4.2 State-equivalent histories violate the margin equality

Reversing the two updates changes answer evidence in both models despite the shared final graph. The prespecified interaction is $\widehat{\theta} = 5.781$ for Qwen3-VL, with a 95% interval of $[3.625, 8.073]$, and $\widehat{\theta} = 3.671$ for Gemma, with an interval of $[2.581, 4.855]$ (Figure 2a). Both multiplicity-adjusted tests support a positive interaction, which refutes the margin equality required by final-state sufficiency. For each model, 21 of 24 matched-instance contributions have the positive orientation fixed before held-out evaluation (Appendix E.1).

The two models reach this result through different condition-level patterns (Figure 2b). For Qwen3-VL, the control-adjusted order contrast is greater for degree than for shortest path under text presentation, but this ordering reverses for images. For Gemma, the degree contrast is greater in both presentations, with a much larger gap in text. Its large negative text shortest-path contrast accounts for most of the positive interaction. Thus, the shared aggregate result reflects presentation-dependent differences between question families, not a uniform direction of history sensitivity across conditions. Exact condition means and their affected/control decomposition appear in Appendix E.2.

4.3 The margin inequality extends beyond K_4 for Gemma

Gemma's history-sensitive margin also appears on larger, non-complete graphs. All seven calibration-eligible panels reject the joint equality of the four control-adjusted order contrasts after Holm correction over all 30 prespecified comparisons; the largest adjusted p-value is 8.9×10^{-4}. These panels contain 224 held-out graph clusters and cover component bridging, cycle closure, and the symmetry-breaking chord at both $n = 6$ and $n = 8$, together with hub attachment at $n = 6$. The result therefore extends beyond the dense K_4 endpoint to four graph families and both tested sizes.

The condition-level estimates show the most consistent pattern for shortest-path questions: all seven image intervals and five of seven text intervals lie below zero. Degree contrasts instead vary in sign across families and presentations. These descriptive differences help explain the joint results without implying a common direction across conditions. The remaining model–family–size combinations did not meet the calibration prerequisites for history inference, so the extension establishes topology scope for Gemma while broader cross-model coverage remains open. Appendices F.4 to F.6 report the condition estimates, eligibility decisions, and scope of inference.

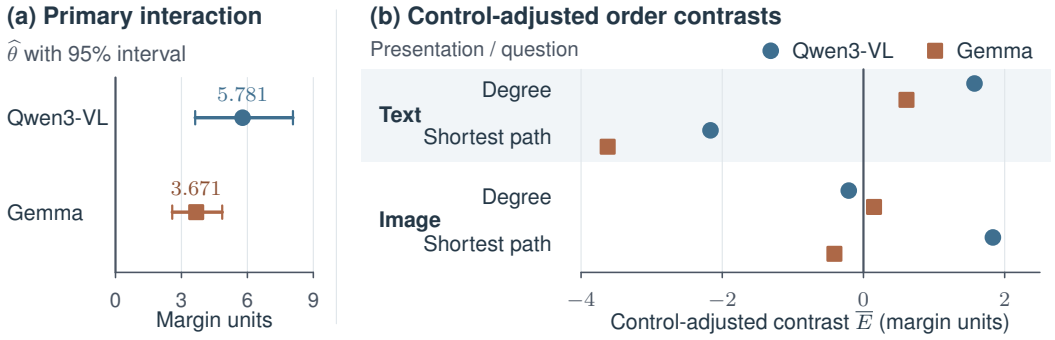

Figure 2: **Primary matched-history test.** (a) The prespecified interaction $\widehat{\theta}$ compares control-adjusted degree-minus-shortest-path contrasts between text and image presentations; final-state sufficiency requires zero. (b) Mean control-adjusted order contrasts $\overline{E}_{M,Q} = \overline{\Delta}^{A}_{M,Q} - \overline{\Delta}^{U}_{M,Q}$, grouped by initial-graph presentation and question family. Each Δ is an $H_{21} - H_{12}$ correct-answer margin difference; A/U denote update-relevant/control questions. Blue circles denote Qwen3-VL and orange squares Gemma. Values summarize 24 matched instances. Bars in (a) are 95% bootstrap intervals; (b) shows means without intervals.

4.4 CORRECT FINAL-STATE DESCRIPTIONS ATTENUATE THE BEHAVIORAL CONTRAST

In the separate one-update diagnostic, a correct final-state description reduces the magnitude of the contrast between graph-changing and state-preserving instructions in three of four model–format comparisons. The attenuation intervals are above zero for Qwen3-VL with either a correct text or image description and for Gemma with a correct image; the corresponding adjusted tests support attenuation (the upper comparisons in Figure 3b). Gemma's correct-text comparison is inconclusive. Explicit endpoint information can therefore reduce this behavioral contrast, with the supported comparisons varying by model and description format.

The signed estimates show that attenuation need not preserve the contrast's sign (Figure 3a). For Gemma, the aggregate changes from 3.459 without a description to -0.719 with a correct image. Although it crosses zero, its absolute magnitude decreases. The prespecified attenuation measure takes absolute values after averaging signed instance contributions, so this aggregate reduction need not occur in every instance. Full estimates and instancewise summaries appear in Appendix E.3.

The remaining contrast also depends on which image is shown. Relative to the correct image, the stale image leaves a larger absolute aggregate by $\widehat{\xi} = 2.142$ for Qwen3-VL and 7.303 for Gemma. Both adjusted tests support this difference (the lower comparison in Figure 3b). This establishes sensitivity to the displayed image condition, rather than isolating graph-state content: the correct and stale renderings also differ in other visual features (Appendices B.3 and E.3).

4.5 INTERNAL EVIDENCE ACROSS LOCATIONS

We return to the core histories to compare relational agreement and causal influence at the two internal locations. Selectivity controls then test whether pre-answer influence depends on the matched history–question pairing.

Relational agreement. The two models organize history differences across question conditions more similarly at the second-update span than at the pre-answer position. The paired second-update-minus-pre-answer difference in agreement is .905 with a 95% interval of $[.774, 1.043]$ for text and .988 with an interval of $[.821, 1.147]$ for images; both adjusted tests support the location difference (Figure 4a). Location-specific correlations are descriptive and are reported in Appendix G.

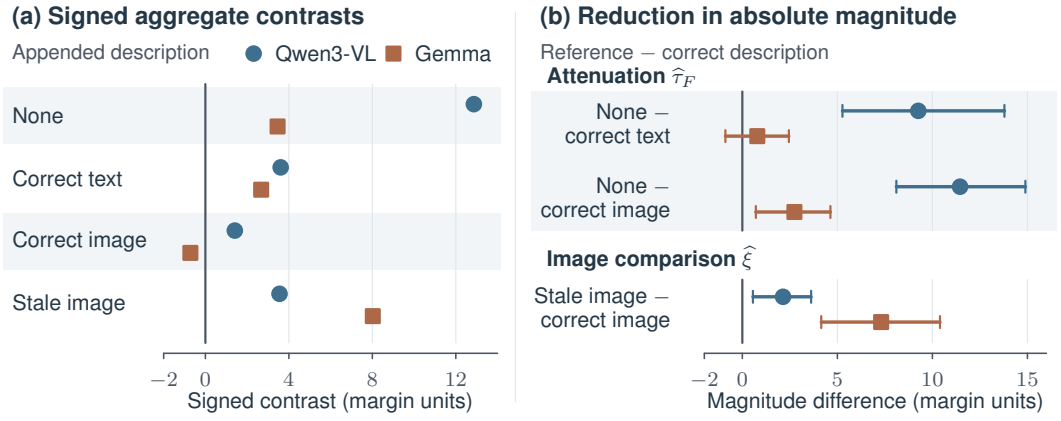

Figure 3: **Final-state description diagnostics.** (a) Signed aggregate contrasts $\widehat{\theta}_c^{\text{final}}$ from state-preserving-minus-graph-changing margins, after control adjustment and the presentation–question-family interaction. Rows name appended descriptions, not initial presentations. (b) Differences of absolute aggregate magnitudes, in the subtraction order shown: no description minus correct text/image ($\widehat{\tau}_F$), and stale minus correct image ($\widehat{\xi}$). Positive values indicate smaller magnitude with a correct description. The stale image depicts the graph-changing condition's initial graph. Blue circles denote Qwen3-VL and orange squares Gemma; each model is shown in a fixed row position. Points summarize 24 matched instances; bars in (b) are 95% bootstrap intervals.

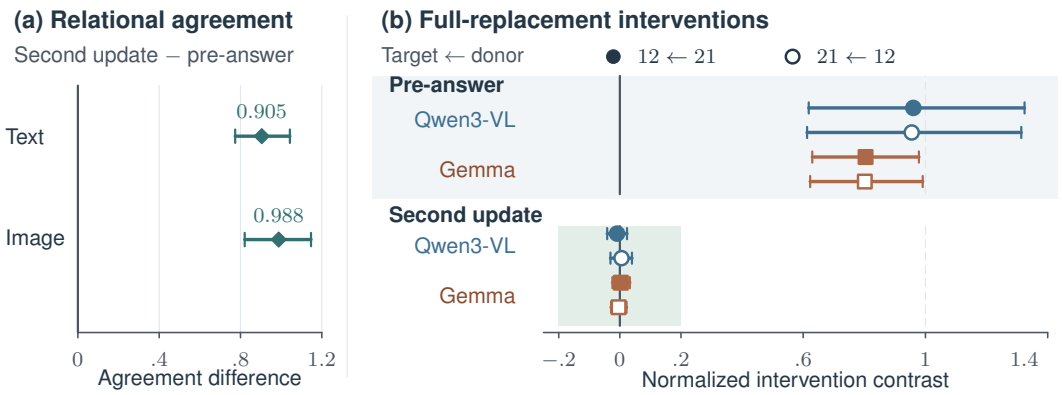

Figure 4: **Internal evidence at two locations.** (a) Paired second-update-minus-pre-answer differences in cross-model Spearman agreement between representational dissimilarity matrices, by initial-graph presentation (teal diamonds). (b) Direction-aligned, control-adjusted full-replacement contrasts divided by each model's observed $\widehat{\theta}$. Blue circles (Qwen3-VL) and orange squares (Gemma) use filled/open markers for the target ← donor directions. The dashed line at one denotes the primary contrast's magnitude; the green $[-.20, .20]$ band applies only to second-update equivalence. Points and 95% bootstrap intervals summarize 24 matched instances.

Causal influence. Activation replacement yields a different location profile. At the second-update span, normalized full-replacement contrasts range from $-.009$ to $.005$. All four 95% intervals lie within the $[-.20, .20]$ equivalence region shaded in Figure 4b, meeting the prespecified small-effect criterion. At the pre-answer position, both replacement directions produce positive intervention contrasts in each model: all four unnormalized intervals are above zero, and at least 21 of 24 instance contributions are positive in each comparison. On the normalized scale, the effects range from .955 to .960 for Qwen3-VL and from .802 to .804 for Gemma, where one denotes the magnitude of the original primary contrast (Figure 4b).

Thus, stronger relational agreement occurs at the update span, while larger intervention contrasts occur before the answer. Intermediate interpolation strengths follow this pattern (Appendix H.1).

Selectivity. The pre-answer influence is not confined to the matched opposite-history donor. Alternative donors retain presentation and question family, testing whether influence extends across graph instances or across questions about the same graph. In both models and both directions, different-question donors produce effects larger in absolute magnitude than matched replacement (Table 13). Qwen3-VL also shows substantial movement on matched-control questions, measured both in absolute terms and through their contribution to the signed interaction, despite small effects from all norm-matched random directions. For Gemma, nine of ten random-direction comparisons exceed the .33 ratio threshold.

Identity replacement has exactly zero effect in every model–direction pair, but at least one small-effect control exceeds its threshold in every pair. The pre-answer state therefore influences the measured contrast without satisfying the prespecified selectivity criterion. The full control matrix and the unavailable cross-family transfer comparison are documented in Appendix H.3.

5 Conclusion

Reaching the same final graph does not guarantee that a model provides the same support for the correct answer. Matched histories reveal update-order dependence in both core models, and the held-out extension shows that Gemma's result spans larger, non-complete graph families. Correct final-state descriptions can reduce a separate update-condition contrast. Internally, cross-model relational agreement is greater at the second update, while activation replacement has greater influence near the answer; the controls do not meet the prespecified selectivity criterion. These findings distinguish successful endpoint answers from reliance on the endpoint alone. Evaluations of dynamic reasoning should test invariance directly across state-equivalent histories.

AI USE STATEMENT

Generative AI models evaluated in this work were used solely as experimental systems, as described in the paper; the disclosure below concerns the use of generative AI tools to assist the research and writing workflow. Generative AI tools were used only for editing and improving software code and for improving the language and readability of the manuscript. They were not used to design the research methodology or experiments, implement the proposed methods, formulate or prove mathematical claims, generate benchmark instances or datasets, interpret experimental results, or develop the scientific conclusions. All AI-assisted code and text were reviewed and verified by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The appendix documents the task construction and prompt templates (Appendix B), model configurations and internal locations (Appendix C), and the complete estimands, inference procedures, intervention definitions, and selectivity controls (Appendix D). It also reports the full behavioral results (Appendix E); the topology-scope construction, eligibility record, held-out tests, and inferential scope (Appendices F.1, F.2, F.4 and F.6); the relational-agreement analysis (Appendix G); the intervention and selectivity results (Appendix H); and the competence-screened discovery, exploratory topology contrasts, and localization studies across graph families and sizes (Appendix I). The development–held-out separation, evaluation counts, and reproduction materials for the core K_4 study and held-out topology extension appear in Appendix J. Together, these sections specify the matched analysis units, aggregation order, resampling procedures, decision rules, model identifiers, and prompt conditions used for every confirmatory claim.

REFERENCES

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923. PMLR, 2016. URL https://proceedings.mlr.press/v48/abel16.html.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 191–203. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4820. URL https://aclanthology.org/W19-4820/.

Qihang Ai, Jiafan Li, Jincheng Dai, Jianwu Zhou, Lemao Liu, Haiyun Jiang, and Shuming Shi. Advancement in graph understanding: A multimodal benchmark and fine-tuning of vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7485–7501, 2024. doi: 10.18653/v1/2024.acl-long.404. URL https://aclanthology.org/2024.acl-long.404/.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Zahra Babaiee, Peyman Kiasari, Daniela Rus, and Radu Grosu. Visual graph arena: Evaluating visual conceptualization of vision and multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2081–2113, 2025. URL https://proceedings.mlr.press/v267/babaiee25a.html.

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

Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu, and Henry Kvinge. Understanding the inner-workings of language models through representation dissimilarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6543–6558, 2023. doi: 10.18653/v1/2023.emnlp-main.403. URL https://aclanthology.org/2023.emnlp-main.403/.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.47. URL https://aclanthology.org/2025.ijcnlp-long.47/.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6596–6620. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24i.html.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1595–1604, 2018. doi: 10.18653/v1/N18-1144. URL https://aclanthology.org/N18-1144/.

Debarati Das, Ishaan Gupta, Jaideep Srivastava, and Dongyeop Kang. Which modality should I use - text, motif, or image? : Understanding graphs with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 503–519, 2024. doi: 10.18653/v1/2024.findings-naacl.34. URL https://aclanthology.org/2024.findings-naacl.34/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*, pp. 43909–43934, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bf72f65f30eedf5d48da6980ee02b589-Abstract-Conference.html.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *International Conference on Learning Representations*, pp. 19337–19359, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3132d0405fabe24b2a7b6cd7ba9de6b5-Abstract-Conference.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4. URL https://doi.org/10.1016/S0004-3702(02)00376-4.

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 *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, 2025. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Aditya Gupta and Greg Durrett. Tracking discrete and continuous entity state for process understanding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pp. 7–12. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1502. URL https://aclanthology.org/W19-1502/.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, pp. 2483–2503, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a6059857ae5c82ea9726ee9282a7145-Abstract-Conference.html.

Michael Hanna, Roberto Zamparelli, and David Mareček. The functional relevance of probed information: A case study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 835–848. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.58. URL https://aclanthology.org/2023.eacl-main.58/.

Romain Harang, Jason Naradowsky, Yaswitha Gujju, and Yusuke Miyao. Tracking world states with language models: State-based evaluation using chess. *arXiv preprint arXiv:2508.19851*, 2025. doi: 10.48550/arXiv.2508.19851. URL https://arxiv.org/abs/2508.19851.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668. Curran Associates, Inc., 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

Lucas Hayne, Heejung Jung, and R. McKell Carter. Does representation similarity capture function similarity? *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=YY2iA0hfia.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3835–3855, 2023. doi: 10.18653/v1/2023.acl-long.213. URL https://aclanthology.org/2023.acl-long.213/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34429–34452, 2025. URL https://proceedings.mlr.press/v267/li25r.html.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, and Min Zhang. VisionGraph: Leveraging large multimodal models for graph theory problems in visual context. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27903–27919, 2024. URL https://proceedings.mlr.press/v235/li24ab.html.

Michael L. Littman and Richard S. Sutton. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, pp. 26486–26515, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. Order-independence without fine tuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72818–72839, 2024. doi: 10.52202/079017-2318. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/85529bc995777a74072ef63c05bedd30-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

Hyunjong Ok and Jaeho Lee. Lost in the prompt order: Revealing the limitations of causal attention in language models. In *Findings of the Association for Computational Linguistics:*

ACL 2026, pp. 38566–38587, 2026. doi: 10.18653/v1/2026.findings-acl.1921. URL https://aclanthology.org/2026.findings-acl.1921/.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Amin Rakhsha, Thomas Hehn, Pietro Mazzaglia, Fabio Valerio Massoli, Arash Behboodi, and Tribhuvanesh Orekondy. LUMINA: Long-horizon understanding for multi-turn interactive agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3913–3926, 2026. doi: 10.18653/v1/2026.findings-acl.190. URL https://aclanthology.org/2026.findings-acl.190/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Kiamehr Rezaee, Jose Camacho-Collados, and Mohammad Taher Pilehvar. Exploring state tracking capabilities of large language models. *arXiv preprint arXiv:2511.10457*, 2025. doi: 10.48550/arXiv.2511.10457. URL https://arxiv.org/abs/2511.10457.

Kyle Richardson, Ronen Tamari, Oren Sultan, Dafna Shahaf, Reut Tsarfaty, and Ashish Sabharwal. Breakpoint transformers for modeling and tracking intermediate beliefs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9703–9719. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.658. URL https://aclanthology.org/2022.emnlp-main.658/.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4902–4918. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.422. URL https://aclanthology.org/2021.findings-emnlp.422/.

Yuanfu Sun, Yuanhang Ren, Kang Li, Chuanhao Ji, Jiaxi Li, Jiajin Liu, Ninghao Liu, and Qiaoyu Tan. GraphVerse: A comprehensive visual graph reasoning benchmark for multimodal large language models. *arXiv preprint arXiv:2608.06769*, 2026. doi: 10.48550/arXiv.2608.06769. URL https://arxiv.org/abs/2608.06769.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6408–6417, 2020. doi: 10.18653/v1/2020.emnlp-main.520. URL https://aclanthology.org/2020.emnlp-main.520/.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26941–26975, 2024. doi: 10.52202/079017-0846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Guanyu Wang, Xu Chu, Zhijie Tan, Xinrong Chen, Tong Mo, and Weiping Li. MuSe: Multi-stage graph reasoning via vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10442–10462, 2026. doi: 10.18653/v1/2026.acl-long.476. URL https://aclanthology.org/2026.acl-long.476/.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *Advances in Neural Information Processing Systems*, volume 36, pp. 30840–30861. Curran Associates, Inc., 2023. doi: 10.52202/075280-1345. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/622afc4edf2824a1b6aaf5afe153fa93-Abstract-Conference.html.

Xueqing Wu, Sha Li, and Heng Ji. OpenPI-C: A better benchmark and stronger baseline for open-vocabulary state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7213–7222, 2023. doi: 10.18653/v1/2023.findings-acl.452. URL https://aclanthology.org/2023.findings-acl.452/.

Takuma Yoneda, Jiading Fang, Peng Li, Huanyu Zhang, Tianchong Jiang, Shengjie Lin, Ben Picker, David Yunis, Hongyuan Mei, and Matthew R. Walter. Statler: State-maintaining language models for embodied reasoning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15083–15091, 2024. doi: 10.1109/ICRA57147.2024.10610634. URL https://doi.org/10.1109/ICRA57147.2024.10610634.

Zike Yuan, Ming Liu, Hui Wang, and Bing Qin. GraCoRe: Benchmarking graph comprehension and complex reasoning in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7925–7948. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.531/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Li Zhang, Hainiu Xu, Abhinav Kommula, Chris Callison-Burch, and Niket Tandon. OpenPI2.0: An improved dataset for entity tracking in texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 166–178, 2024a. doi: 10.18653/v1/2024.eacl-long.10. URL https://aclanthology.org/2024.eacl-long.10/.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, and Yueting Zhuang. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19228–19252, 2024b. doi: 10.18653/v1/2024.emnlp-main.1072. URL https://aclanthology.org/2024.emnlp-main.1072/.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13603–13621, 2025. doi: 10.18653/v1/2025.acl-long.668. URL https://aclanthology.org/2025.acl-long.668/.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, pp. 19426–19454, 2024. URL https://proceedings.iclr.cc/paper_files/paper/

2024/hash/54dd9e0cff6d9214e20d97eb2a3bae49-Abstract-Conference.html.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30678–30701, 2025. doi: 10.18653/v1/2025.acl-long.1482. URL https://aclanthology.org/2025.acl-long.1482/.

A Extended Related Work

A.1 Tracking states in changing worlds

State-tracking benchmarks often present an initial world and a sequence of changes. ProPara annotates entity location and existence throughout natural process descriptions (Dalvi et al., 2018). OpenPI extends the task to open-vocabulary entities, attributes, and state changes (Tandon et al., 2020). OpenPI-C cleans the annotations, revises evaluation, and models temporal dependence, while OpenPI2.0 canonicalizes entities and attributes and adds salience annotations (Wu et al., 2023; Zhang et al., 2024a). Structured models for ProPara combine continuous entity representations with globally constrained discrete state transitions (Gupta & Durrett, 2019), while TRIP couples an endpoint plausibility judgment to annotated physical states and supporting conflicts (Storks et al., 2021). These benchmarks ask a model to recover or justify intermediate and final states, usually through state labels, structured changes, or supporting evidence.

Recent work examines state tracking in pretrained language models under more controlled conditions. Entity-tracking tasks test whether models infer final entity states after sequences of object placement, removal, and movement rather than rely on lexical cues (Kim & Schuster, 2023). Synthetic evaluations vary positions, object exchanges, and binary state changes (Rezaee et al., 2025). Permutation-composition tasks show that models can learn distinct state-tracking algorithms with different generalization properties (Li et al., 2025). Breakpoint modeling directly supervises queryable beliefs at intermediate positions in a sequence (Richardson et al., 2022), while LUMINA uses oracle counterfactuals to separate state tracking from planning and other demands of long-horizon interaction (Rakhsha et al., 2026). These studies evaluate state recovery or the computations that support it. Our question begins after two histories are known to reach the same state: does the answer distribution still distinguish the histories?

A.2 Final-state information and coherent world models

Some methods make the current state explicit rather than requiring a model to recover it only from the full history. Statler maintains a textual estimate of the world state and conditions later embodied decisions on that estimate (Yoneda et al., 2024). Our auxiliary experiment does not maintain a state; it measures how correct or stale final-state descriptions change a separate graph-changing-versus-state-preserving contrast. This role differs from both a state-maintenance method and the primary reversed-order test.

A complementary line evaluates whether model predictions imply a coherent state space. Vafa et al. formalize world-model recovery for deterministic finite automata and introduce Myhill–Nerode-inspired criteria based on state equivalence and distinguishability, with applications to games, logic puzzles, and navigation (Vafa et al., 2024). Harang et al. evaluate predicted chess states through the distributions of legal moves they afford (Harang et al., 2025). Representation studies also recover linear spatial and temporal variables from language-model activations (Gurnee & Tegmark, 2024). These approaches motivate a distinction between recovering state information and using it as the sole basis for a current-state answer. Our construction starts from a known symbolic endpoint and tests the latter requirement through matched histories; it does not infer an unknown world model from trajectories.

A.3 State sufficiency and history abstraction

Classical sequential models define a state by the information from history that must be retained for prediction or control. Predictive state representations summarize history through predictions of future action–observation sequences (Littman & Sutton, 2001). Work on Markov decision processes instead studies when states may be grouped while preserving decision-relevant dynamics or behavior, from exact equivalence and model minimization (Givan et al., 2003) to approximate abstractions with bounded behavioral loss (Abel et al., 2016). In each case, state is a history summary whose adequacy is defined by what it preserves.

Our setting does not learn a transition model, policy, or abstract state. The endpoint is a known symbolic graph, and final-state sufficiency is a task-specific factorization claim: once that endpoint is fixed, the answer distribution for a fixed current-state question should no longer depend on the

path used to reach it. Matched state-equivalent histories make a violation of this claim observable even when the most likely answer remains unchanged.

A.4 Graphs across visual and textual representations

Graph benchmarks introduce a distinct challenge because a model must recover relational structure before reasoning about it. Textual evaluations formulate graph algorithms as natural-language problems, compare alternative graph encodings, and expand the range of graph structures and reasoning tasks (Wang et al., 2023; Fatemi et al., 2024; Yuan et al., 2025). These studies show that task performance depends on how a graph is serialized as well as on its topology. Visual benchmarks add a perception problem: VisionGraph evaluates connectivity and shortest paths from rendered graphs (Li et al., 2024); other work evaluates diverse graph formats or compares text, motif, and image encodings (Ai et al., 2024; Das et al., 2024). VGCure, Visual Graph Arena, and GraphVerse broaden the task, topology, and layout coverage of visual graph reasoning (Zhu et al., 2025; Babaiee et al., 2025; Sun et al., 2026).

Related methods use relation graphs as one class of abstract image in synthetic instruction data (Zhang et al., 2024b) or render task-relevant subgraphs to make larger inputs tractable (Wang et al., 2026). These studies establish that modality, layout, and visual complexity can affect measured graph-reasoning performance. Our matched diagnostic fixes the endpoint graph, question, answer set, node labels, and presentation within each pair; only the order of two valid state transitions changes. The initial graph may be presented as text or an image, but presentation is a controlled factor rather than evidence that either modality is generally superior.

A.5 Prompt order and state-equivalent histories

Language-model predictions vary with the order of demonstrations, premises, answer options, and prompt blocks (Lu et al., 2022; Chen et al., 2024; Pezeshkpour & Hruschka, 2024; Zheng et al., 2024; Ok & Lee, 2026). Long-context performance also varies with the position of task-relevant evidence (Liu et al., 2024). These results establish that static prompt organization can affect a model even when task-relevant content is preserved. McIlroy-Young et al. instead remove order information for designated subsequences to enforce invariance to permutations within those sets (McIlroy-Young et al., 2024). Their objective is constructive; ours is to diagnose an existing model without changing its input representation.

Our order manipulation differs in what the order denotes. State-equivalent histories are ordinary transition sequences that pass through different intermediate graphs, although they contain the same updates and end at the same graph. Reversing them therefore changes the trajectory through the world while fixing the endpoint, question, answer set, node labels, and answer order. This is the distinction between our test and a permutation of static prompt content.

A.6 Internal representations of changing states

Mechanistic studies ask what state information is represented and whether later computation uses it. Work on Othello-GPT decodes an internal board state and uses interventions to test whether it controls move predictions (Li et al., 2023). Controlled finite-state and permutation tasks identify candidate computations for state transitions through probing and activation replacement (Zhang et al., 2025; Li et al., 2025). Propositional probes further show that internal representations can retain state information not reflected in an overt answer (Feng et al., 2025).

Probe performance establishes extractability, not task use. Control tasks test whether accuracy reflects information in the representation rather than probe capacity (Hewitt & Liang, 2019), and task-relevance studies show that decodable information can be present even when the trained computation does not require it (Ravichander et al., 2021). Causal interventions can narrow this gap by identifying the locations at which encoded information affects behavior (Hanna et al., 2023). Our relational-agreement analysis asks whether two models organize a fixed set of matched history contrasts similarly; activation replacement separately tests whether the selected states affect the behavioral contrast.

A.7 Relational agreement

Representational similarity analysis compares pairwise relations among conditions and therefore does not require individual units in two systems to align (Kriegeskorte et al., 2008). Canonical-correlation methods and centered kernel alignment differ in the transformations of representation space they treat as irrelevant (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019); related comparisons have been applied to language-model representations (Abnar et al., 2019; Brown et al., 2023). Agreement alone need not entail functional similarity or show that later computation uses the shared organization (Hayne et al., 2024). Our cross-model relational agreement concerns the ordering of pairwise distances. We use a paired comparison between locations, with the individual correlations providing descriptive context (Appendices D.4 and G).

A.8 Causal influence and selectivity

Causal methods change an internal variable and measure the behavioral consequence. Causal-abstraction analyses align neural representations with high-level variables and validate those alignments through interchange interventions (Geiger et al., 2021; 2022; 2024). Neighboring approaches include causal mediation (Vig et al., 2020), removal of decodable information (Elazar et al., 2021), causal tracing (Meng et al., 2022), and corruption–intervention analyses for vision–language models (Golovanevsky et al., 2025).

An observed effect remains conditional on the intervention and outcome. CausalGym evaluates whether interpretability methods produce intended behavioral changes (Arora et al., 2024), while reliability analyses of causal probes distinguish completeness from preservation of non-target properties (Canby et al., 2025). Localization and intervention conclusions can also diverge across editing objectives, behavioral metrics, or corruption schemes (Hase et al., 2023; Zhang & Nanda, 2024); a replacement may even activate a dormant pathway that the unmodified model does not normally use (Makelov et al., 2024).

Our matched opposite-history replacement therefore supports *causal influence* only for the specified intervention site and correct-answer margin. *Selectivity* is the stronger claim that this influence is specific to the relevant history–question relation rather than a broad change in answer preferences. Our controls test both limited nonspecific disturbance and transfer of the relevant influence across question families. Their definitions and decision rules appear in Appendix D.6.

B Task Construction and Design Accounting

B.1 Primary state-equivalent update histories

Each primary matched graph instance begins with the complete graph on four nodes, K_4, after removing two edges that share an endpoint. The two update histories add those edges in opposite orders. Both requested edges are absent when added, so every instruction changes the graph and both update histories reach K_4. The identifiers e_1 and e_2 denote roles within an instance; they do not privilege a fixed node label or image position. Across the 24 held-out instances, the topology is fixed while node labels, question realizations, and image coordinates vary deterministically with the instance construction.

The primary question battery comprises two matched update-relevant/control pairs, one from the degree family and one from the shortest-path family. The two questions in each pair belong to the same question family and are matched by construction. The answer to the update-relevant question changes during the two updates, whereas the answer to the matched-control question does not. Crossing the four questions with two update histories and two initial-graph presentations yields $24 \times 4 \times 2 \times 2 = 384$ held-out prompt conditions per model.

Text inputs describe the initial graph with an undirected adjacency matrix. Image inputs depict the same graph; the two textual updates and the question then follow the image. One text condition is:

```
Initial graph.  Node order:  A C D M
Adjacency matrix (undirected; rows and columns use the node
order):
                  A C D M
A | 0 1 0 1
C | 1 0 1 1
D | 0 1 0 0
M | 1 1 0 0
Update 1 of 2:  Add the edge between D and A.
Update 2 of 2:  Add the edge between D and M.
What is the degree of D?
```

In this example, the update-relevant degree question concerns D, whose degree changes from one to three; its control concerns C, whose degree remains three. The update-relevant shortest-path question concerns D and A, whose distance changes from two to one; its control concerns A and M, whose distance remains one. Both histories receive these same four questions after the updates. The controls therefore share the graph and update sequence with the relevant questions while asking about properties that neither addition changes.

The paired history exchanges only the two update sentences. For image presentation, node coordinates vary independently of the canonical graph roles. A fixed spatial location therefore cannot identify an update-relevant node or relation across instances. Text and image conditions retain the same graph identity, node labels, history, question identity, and valid answer set. Image coordinates are presentation features and are not treated as matched analysis variables.

B.2 Held-out topology-scope construction

The topology extension retains the reversed additions and matched questions while replacing the K_4 endpoint with non-complete graphs. Five families at six and eight nodes give ten fixed templates, each with 32 held-out label and layout variants. Appendix F.1 defines the initial graphs, update edges, questions, and split construction; Appendix F.2 specifies the independent calibration checks.

B.3 Final-state-description design

The separate final-state design uses two one-update conditions. In the graph-changing condition, the initial graph is $K_4 \setminus \{e^\star\}$ and the instruction adds $e^\star$. In the state-preserving condition, the initial graph is K_4 and the same instruction requests an edge that is already present. Both conditions reach K_4 and use the same questions and correct answers. Their initial graphs differ, and the instruction changes the graph in only one condition. This design therefore does not isolate update order under a shared initial graph, as H1 does, even though the endpoints are state-equivalent.

Each instance has a six-question battery: two matched update-relevant–control degree pairs and one matched update-relevant–control shortest-path pair. Analysis first subtracts the matched-control contrast from the update-relevant contrast within each pair and then averages the two degree matches. The degree and shortest-path families therefore receive equal weight in the presentation–question-family interaction despite containing different numbers of query pairs.

The description is inserted after the update and before the question. Its format is independent of the initial-graph presentation: either a text or an image initial graph can be followed by either description format. H2/H3 use four conditions: no description, correct text, correct image, and stale image. The larger archive crosses two description formats with five content conditions and separately includes the no-description reference. Table 1 distinguishes the conditions used in H2/H3 from the additional archived diagnostics.

The stale content is fixed to $K_4 \setminus \{e^\star\}$ in both update conditions. It is therefore the graph-changing condition's pre-update state but not the state-preserving condition's initial graph. The correct and stale image conditions are complete renderings whose layouts and other visual features are not matched feature by feature. Their H3 contrast therefore measures sensitivity to the displayed image

Table 1: Post-update content conditions and their roles. Each of the five description contents is rendered in text and image form, independently of the initial presentation. H2/H3 use only the listed references and correct/stale conditions; the other archived contents do not enter their estimands.

Content	Description	Role in this paper
No description	Optional slot omitted	H2 reference c_0
Correct final	Full description of K_4	Text and image H2 conditions c_T, c_I; image H3 reference
Stale initial	Full description of the five-edge graph $K_4 \setminus \{e^\star\}$	Image H3 condition c_S
Matched wrong	Five-edge graph containing $e^\star$ but omitting another edge	Archived content diagnostic
Surface-reordered correct	Complete K_4 with a changed text order or image layout	Archived surface diagnostic
Target edge only	Explicitly incomplete reminder that $e^\star$ is present	Archived partial-state diagnostic

condition: it combines the intended graph-state difference with any remaining visual differences between the renderings.

The held-out no-description design contains $24 \times 2 \times 2 \times 6 = 576$ conditions per model: 24 graph instances, two update conditions, two initial-graph presentations, and six questions. The description archive contains $24 \times 2 \times 2 \times 2 \times 5 \times 6 = 5{,}760$ conditions per model after crossing the same factors with two description formats and five contents. H2 uses c_0 and the correct-final content in each format. H3 uses only the correct-final and stale-initial image conditions. The remaining archive conditions do not enter the H2 or H3 estimands.

B.4 Prompt content and valid answers

Each construction uses prompt wording that was fixed before held-out evaluation. The logical slots and their order are summarized in Table 2.

Table 2: Prompt and response slots. "Absent" means that the optional slot is omitted rather than replaced with padding.

Slot	Primary construction	Final-state construction
Initial graph	Text adjacency matrix or graph image	Text adjacency matrix or graph image
Update	Two ordered textual edge additions	One textual edge-addition instruction
Final-state description	Absent	Absent, or a text or image condition from Table 1
Question	Degree or shortest-path query	Six-query matched battery
Response	One valid complete answer and model terminator	One valid complete answer and model terminator

The final-state prompt follows the template below. Bracketed text denotes the condition-specific graph or label rather than an additional instruction:

```
Graph before applying the update:
[initial adjacency matrix or graph image]

Add the edge between [node 1] and [node 2]. All other edges
remain unchanged. The graph is simple, so an edge already
in the requested state remains unchanged. Answer about the
resulting graph.

[optional post-update description]

[question]
```

For every complete post-update description, the optional block begins with "The following terminal evidence is authoritative for the question. Use its graph content as the current state." The corresponding adjacency matrix or image follows this sentence. For the target-edge-only condition, it

instead begins with "The following is a target-edge reminder only, not a complete graph. It states one current edge fact." The no-description condition omits the entire block, including its label.

The degree-question template in the final-state construction is "How many direct neighbors does node [node] have? Answer with one integer." The shortest-path template is "What is the shortest-path length from node [source] to node [target], counted in edges? If no path exists, answer unreachable." The primary templates use "What is the degree of [node]?" and "What is the shortest-path distance between [source] and [target]?" as illustrated above.

In an image condition, the image occupies the initial-graph slot while the updates and question remain textual. In the final-state design, the optional description appears after the update and before the question. Each checkpoint's native conversation template formats this common content and opens the assistant response. Model-specific role markers and control tokens surround, but do not replace or augment, the semantic blocks specified here. The immutable checkpoint revisions and answer terminators are reported in Appendix C.1.

For the core K_4 and final-state constructions, candidate scoring uses the full valid-answer set for each question on a four-node simple graph. Degree candidates are $\{0, 1, 2, 3\}$, whereas shortest-path candidates are $\{1, 2, 3, \texttt{unreachable}\}$. Thus shortest-path responses are not all integers. In the topology-scope extension with n nodes, the corresponding supports are $\{0, \ldots, n-1\}$ for degree and $\{1, \ldots, n-1, \texttt{unreachable}\}$ for shortest path; edge-adoption diagnostics use `yes` and `no`. Each scored candidate contains one complete answer followed by the checkpoint-specific end-of-turn token, with no explanatory continuation. Token-level details and the distinction between likelihood scoring and diagnostic generation appear in Appendix C.1.

C Models, Scoring, and Internal Locations

C.1 Evaluation setting

The core behavioral and internal analyses use Qwen3-VL-8B-Instruct and Gemma-4-12B-it. The topology extension adds Qwen3.5-9B for behavioral evaluation, with no activation analysis for that checkpoint. All reported margins use complete-answer likelihoods rather than freely generated answers.

Checkpoints. The core study evaluates `Qwen/Qwen3-VL-8B-Instruct`, ModelScope revision `5d854aab08710c16b980ec6d603d863b3821b915`, and `google/gemma-4-12B-it`, revision `657684fef0b5ac5d6bff39284ceb6ec3710b700e`. All weight shards were verified against recorded SHA-256 digests. These immutable revisions are used for the core held-out behavioral measurements, activation capture, and interventions.

The held-out topology-scope extension additionally evaluates `Qwen/Qwen3.5-9B` (Qwen Team, 2026), revision `e138e55eb2e1f1c59ebbf0f815bb972649fc22c6`. Qwen3.5 participates in the extension's calibration, development, and held-out behavioral splits; it does not enter the core K_4 estimands or any activation analysis. Exact checkpoint roles and transfer provenance are reported in Appendices F.1 and F.5.

Complete-answer scoring. We score each valid answer by teacher-forcing the complete assistant response, including the checkpoint-specific, tokenizer-verified end-of-turn token. Qwen3-VL uses `<|im_end|>` (token ID 151645), Qwen3.5 uses the same token string with token ID 248046, and Gemma uses `<turn|>` (token ID 106). The score is the sum of the conditional token log-likelihoods over the answer and its terminator, without length normalization. An alternative such as `unreachable` is therefore scored as a complete response; if it spans multiple tokens, every token contributes. No response is sampled for the reported correct-answer margins, so sampling temperature and top-p do not apply.

Generation and runtime. The core evaluation configuration specifies greedy decoding with at most 32 new tokens when generation is required. Qwen3-VL's optional reasoning mode is disabled with `enable_thinking=false`. The primary analyses use complete-answer scoring rather than generated responses. The preliminary competence summaries in Appendix I.5 come from a separate localization evaluation.

The topology extension uses BF16 weights without quantization and the `eager` attention implementation. Diagnostic generation is greedy (one beam, no sampling), with reasoning disabled and a maximum of 32 new tokens. Candidate scoring is performed separately for every row. Its calibration accuracy measures and free-generation parse check have different roles, specified in Appendix F.2; generated-answer accuracy is not substituted for candidate accuracy in the eligibility rules.

The frozen core-study environment uses PyTorch 2.11.0, torchvision 0.26.0, Transformers 5.13.0, and safetensors 0.8.0. Inference uses BF16 weights on NVIDIA RTX PRO 6000 Blackwell GPUs, with one visible device assigned to each model run. The core held-out evaluation required approximately 2.4 wall-clock hours.

C.2 Residual-stream locations and token roles

The frozen sites lie at approximately 80% of decoder depth; the exact zero-based layer indices are given below. Measurements use decoder-block outputs at two token roles. The *second-update span* comprises the tokens that express the second textual update. The *pre-answer position* is the final prefix token immediately before the assistant answer begins.

Table 3: Frozen residual-stream sites. Layer indices begin at zero; width is the residual-stream dimension.

Model	Layers	Layer	Width	Attention
Qwen3-VL	36	28	4096	full
Gemma	48	38	3840	sliding

The Qwen3-VL module is `model.language_model.layers.28`, and the Gemma module is `model.language_model.layers.38`. Both locations are decoder-block outputs. Token roles are resolved independently with each model's processor; positions are not transferred across architectures.

The relational-agreement analysis mean-pools the captured activation tensor across the token axis, leaving one hidden vector for each second-update span and one for the single-token pre-answer position. The intervention analysis does not use this pooled vector. It interpolates or replaces the full activation tensor at all selected token positions within a model. Cross-model comparison uses ranks of within-model RDM distances, so the models' different hidden widths do not require explicit coordinate alignment.

Matched opposite-history interventions and selectivity comparisons use the frozen compatibility rules: decoder layer, captured tensor shape, number of selected tokens, and their exact absolute positions must match. The selected token count is not the full prompt length. All selectivity controls use the single-token pre-answer position; matching its absolute position also fixes the prefix length. The implications of these rules for the cross-family reference pool are specified in Appendix D.6.

D Detailed Estimands, Controls, and Inferential Assumptions

This section expands the estimands, intervention controls, and inferential assumptions used for the core K_4 study in Section 3. Each core analysis first reduces matched conditions within one graph instance; aggregation and resampling occur only after this reduction. The core study's 24 held-out instances are a fixed constructed set, so the reported quantities are finite-set summaries rather than expectations over a superpopulation. The topology-scope extension, whose design was fixed before held-out evaluation, uses its own matched clusters, eligibility rules, and joint statistics, summarized below and specified fully in Appendix F.

D.1 Behavioral estimands

Primary history-order estimand. Let $n = 24$ and let $i \in \{1, \dots, n\}$ index a held-out matched graph instance, $M \in \{T, I\}$ a text or image presentation, $Q \in \{d, \mathrm{sp}\}$ a degree or shortest-path question, and $r \in \{A, U\}$ an update-relevant or matched-control question. All quantities are computed

separately for each model. For any four quantities $X_{M,Q}$, define the prespecified presentation–question-family contrast

$$\mathcal{C}(X) = (X_{T,d} - X_{T,\mathrm{sp}}) - (X_{I,d} - X_{I,\mathrm{sp}}). \tag{7}$$

Using the correct-answer margin in Eq. (2), write $m^r_{i,M,Q}(H)$ for the margin under one matched history. The signed order contrast follows the $H_{21} - H_{12}$ orientation:

$$\Delta^r_{i,M,Q} = m^r_{i,M,Q}(H_{21}) - m^r_{i,M,Q}(H_{12}).$$

Subtracting the matched-control contrast yields the control-adjusted order contrast for the update-relevant property,

$$E_{i,M,Q} = \Delta^A_{i,M,Q} - \Delta^U_{i,M,Q}, \tag{8}$$

and the matched-instance contribution to the primary aggregate is

$$\Psi_i = \mathcal{C}(E_i). \tag{9}$$

The corresponding finite held-out-set mean is

$$\widehat{\theta} = \frac{1}{24}\sum_{i=1}^{24} \Psi_i.$$

For the cell-level decomposition, let $\overline{E}_{M,Q} = 24^{-1}\sum_i E_{i,M,Q}$; then $\widehat{\theta} = \mathcal{C}(\overline{E})$. The matched-history equality implied by final-state sufficiency sets every tested order contrast in Eq. (3) to zero and thus implies $\widehat{\theta} = 0$. **H1** is the prespecified directional claim $\widehat{\theta} > 0$. A nonzero finite-set value is incompatible with that equality in the tested construction, whereas a zero aggregate would not establish final-state sufficiency because nonzero cell contrasts could cancel. Inferential support for the directional claim additionally requires the decision rule below.

Role of the crossed contrast. The three comparisons address different aspects of history dependence. The raw contrast $\Delta^r_{i,M,Q}$ asks whether reversing the updates changes the margin for a fixed question. The difference $E_{i,M,Q}$ places that change in relation to the change for an unaffected property under the same histories. It does not assume that the control is itself invariant. The final contrast $\mathcal{C}$ asks whether the difference between degree and shortest-path sensitivity changes with graph presentation. Degree questions concern local adjacency and shortest-path questions concern paths between nodes; text and images provide two presentations of the same graph. Their interaction therefore describes how the queried relation and its presentation jointly shape control-adjusted history sensitivity.

To see what this contrast retains, suppose the cell means can be written as $\overline{E}_{M,Q} = a_M + b_Q$, where a_M depends only on presentation and b_Q only on question family. Then

$$\mathcal{C}(\overline{E}) = \big[(a_T + b_d) - (a_T + b_{\mathrm{sp}})\big] - \big[(a_I + b_d) - (a_I + b_{\mathrm{sp}})\big] = 0.$$

Thus a common shift or additive presentation and question-family effects do not contribute to the interaction. This is an algebraic property of the contrast, not evidence that all sources of order sensitivity have been removed. It also explains why the signed components need not agree: their coefficients encode a difference of differences, not an average tendency across tasks.

A joint test of the raw order-contrast vector asks whether any component departs from zero. The core interaction instead tests one prespecified linear contrast of that vector. It characterizes a particular pattern of departure from final-state sufficiency rather than testing every possible departure. The topology extension uses a separately specified joint statistic Q_E for the four control-adjusted contrasts (Appendices D.3 and F.3). Its joint decisions are distinct from the core study's directional H1 decisions. The two analyses provide complementary evidence without making their statistics interchangeable.

Final-state-description estimands. This diagnostic asks whether providing the reached graph reduces the contrast between a graph-changing instruction and a state-preserving one. It compares different initial graphs with the same final graph, not two orders of the same updates. The two update conditions are denoted $u \in \{\mathrm{chg}, \mathrm{pres}\}$. Let $c \in \{c_0, c_T, c_I, c_S\}$ denote no additional final-state description, a correct text description, a correct image, or a stale image. With $j \in \mathcal{J}_Q$ indexing a matched pair of update-relevant and control questions, the margin is $m^{r,u}_{i,c,M,Q,j}$. There are two

degree pairs and one shortest-path pair: $|\mathcal{J}_d| = 2$ and $|\mathcal{J}_{\text{sp}}| = 1$. We first take the state-preserving minus graph-changing difference within a matched query pair, then average pairs within each question family before applying the presentation–question-family interaction:

$$\Delta^{\text{final},r}_{i,c,M,Q,j} = m^{r,\text{pres}}_{i,c,M,Q,j} - m^{r,\text{chg}}_{i,c,M,Q,j}, \tag{10}$$

$$E^{\text{final}}_{i,c,M,Q} = \frac{1}{|\mathcal{J}_Q|} \sum_{j \in \mathcal{J}_Q} \left(\Delta^{\text{final},A}_{i,c,M,Q,j} - \Delta^{\text{final},U}_{i,c,M,Q,j} \right), \tag{11}$$

$$\Psi^{\text{final}}_{i,c} = \mathcal{C}(E^{\text{final}}_{i,c}), \tag{12}$$

$$\widehat{\theta}^{\text{final}}_{c} = \frac{1}{24} \sum_{i=1}^{24} \Psi^{\text{final}}_{i,c}. \tag{13}$$

For $F \in \{T, I\}$, the reduction produced by a correct final-state description is

$$\widehat{\tau}_F = |\widehat{\theta}^{\text{final}}_{c_0}| - |\widehat{\theta}^{\text{final}}_{c_F}|.$$

H2 tests $\widehat{\tau}_F > 0$ separately for text and image descriptions within each model. The stale-versus-correct image contrast is

$$\widehat{\xi} = |\widehat{\theta}^{\text{final}}_{c_S}| - |\widehat{\theta}^{\text{final}}_{c_I}|.$$

H3 tests $\widehat{\xi} > 0$ within each model. Both summaries take absolute values only after the signed matched-instance contrasts have been averaged. Reversing the common update-condition orientation therefore leaves H2 and H3 unchanged. The presentation index M refers to the initial graph, whereas c specifies the description added after the update. Averaging matched query pairs within Q prevents the two degree pairs from receiving twice the weight of the single shortest-path pair; the subsequent operator $\mathcal{C}$ remains a signed interaction, not an average across question families.

D.2 Behavioral inference

The 95% confidence intervals are percentile intervals from 10,000 bootstrap samples of the matched graph instances, using master seed 20260901. Each draw recomputes the complete contrast from the sampled instances; prompts that share an instance are never resampled independently. These intervals summarize sensitivity to the composition of the fixed constructed set. They do not by themselves license superpopulation generalization.

H1. The positive orientation of Ψ_i was selected on development examples and fixed before held-out evaluation. For each model, H1 tests $\widehat{\theta} > 0$ with a one-sided exhaustive sign-flip calculation. Conditional on the observed magnitudes $|\Psi_1|, \ldots, |\Psi_n|$, it enumerates all 2^n sign configurations. Exactness requires independent sign symmetry across matched instances under the null. Edge roles were fixed rather than randomized, so this symmetry is an inferential assumption.

H2 and H3. For H2, each matched graph instance contributes the two signed contrasts compared in $\widehat{\tau}_F$; for H3 it contributes the contrasts compared in $\widehat{\xi}$. A bootstrap draw resamples these paired values jointly and recomputes the difference between the absolute aggregate means. The one-sided paired-swap calculation enumerates all 2^n combinations of within-instance label exchanges. Its exact interpretation requires the two labels to be exchangeable under the corresponding sharp null of no condition-specific contrast. The final-state conditions were systematically crossed rather than randomly assigned, so label exchangeability is an inferential assumption rather than a consequence of treatment randomization.

Holm correction is applied separately to four prespecified families: the two model-specific H1 tests; the four H2 tests formed by two models and two final-state formats; the two model-specific H3 tests; and the two presentation-specific relational-agreement tests defined below. A behavioral comparison supports its one-sided hypothesis only when its cluster-bootstrap lower bound is positive and its Holm-adjusted test rejects at the stated familywise level. All families use familywise error rate $\alpha = .05$.

D.3 INFERENCE IN THE TOPOLOGY-SCOPE EXTENSION

The topology extension uses the joint control-adjusted statistic Q_E and the supporting update-relevant statistic Q_A, rather than the core interaction $\widehat{\theta}$. These statistics allow condition contrasts of either sign and use the extension's own graph clusters. Appendices F.2, F.3 and F.6 give the calibration criteria, complete inference procedure, and nested cross-model breadth rule.

D.4 RELATIONAL-AGREEMENT ESTIMAND

We compare how the two models organize history differences across questions, not whether their individual hidden coordinates match. Within one graph instance and presentation, each model supplies four difference vectors: degree and shortest path, each with an update-relevant and a control question. The six pairwise distances among those vectors define a relational structure that can be compared across models with different hidden widths.

Let $g \in \{1, 2\}$ index the two models and let $\ell \in \{\mathrm{upd}, \mathrm{ans}\}$ denote the second-update span or pre-answer position. Let $k = (Q, r)$ index the four conditions in $\mathcal{K} = \{(d, A), (d, U), (\mathrm{sp}, A), (\mathrm{sp}, U)\}$, and let d_g be model g's hidden width. After removing the singleton batch axis, write the captured residual-stream tensor as $Z^{(g)}_{i,M,\ell,k}(H) \in \mathbb{R}^{t \times d_g}$, where t is the number of selected token positions, not the full prompt length. It may depend on the prompt and location. Relational agreement uses the mean-pooled vector

$$\overline{\mathbf{z}}^{(g)}_{i,M,\ell,k}(H) = \frac{1}{t} \sum_{p=1}^{t} Z^{(g)}_{i,M,\ell,k,p}(H).$$

This pooling is specific to the relational-agreement analysis. The intervention defined below operates on the complete $t \times d_g$ tensor instead. Exact token roles and residual-stream sites are specified in Appendix C.2. For each condition, define the within-model order-contrast vector

$$\mathbf{v}^{(g)}_{i,M,\ell,k} = \overline{\mathbf{z}}^{(g)}_{i,M,\ell,k}(H_{21}) - \overline{\mathbf{z}}^{(g)}_{i,M,\ell,k}(H_{12}). \tag{14}$$

For distinct $k, k' \in \mathcal{K}$, the RDM entry is the cosine dissimilarity

$$D^{(g)}_{i,M,\ell}(k, k') = 1 - \frac{\langle \mathbf{v}^{(g)}_{i,M,\ell,k}, \mathbf{v}^{(g)}_{i,M,\ell,k'} \rangle}{\|\mathbf{v}^{(g)}_{i,M,\ell,k}\|_2 \|\mathbf{v}^{(g)}_{i,M,\ell,k'}\|_2}.$$

Let $\mathbf{d}^{(g)}_{i,M,\ell}$ contain the six off-diagonal upper-triangular entries of $D^{(g)}_{i,M,\ell}$. Cross-model relational agreement is

$$\rho_{i,M,\ell} = \mathrm{Spearman}\left(\mathbf{d}^{(1)}_{i,M,\ell}, \mathbf{d}^{(2)}_{i,M,\ell}\right).$$

This comparison uses within-model distances and does not align individual representation coordinates across models. Reversing the common orientation of all four order-contrast vectors leaves the cosine dissimilarities unchanged. Relational agreement concerns only the ordering of these pairwise dissimilarities. It does not identify shared coordinates, individual features, neural mechanisms, or algorithms.

Let $J_{i,M} = 1$ when both models have finite, nonzero order-contrast vectors and nonconstant RDM rank vectors at both locations. With $\mathcal{I}_M = \{i : J_{i,M} = 1\}$ and $n_M = |\mathcal{I}_M|$, the finite-set summary for presentation M is

$$\widehat{\zeta}_M = \frac{1}{n_M} \sum_{i \in \mathcal{I}_M} \left(\rho_{i,M,\mathrm{upd}} - \rho_{i,M,\mathrm{ans}}\right). \tag{15}$$

Inference is performed only when at least 18 complete matched graph instances remain, with no imputation. For each presentation, the one-sided alternative is $\widehat{\zeta}_M > 0$. The exhaustive sign-flip calculation assumes that all sign configurations of the matched-instance location differences are equally likely under the null conditional on their magnitudes. Text and image form the two-test Holm family listed above.

D.5 Intervention estimand and normalization

Each target–donor pair shares the matched graph instance, presentation, question, question role, and internal location; only update order differs. The two directions are $12 \leftarrow 21$ and $21 \leftarrow 12$, where the left history is the target and the right history supplies the donor. If the selected residual-stream tensors have shape $t \times d$, interpolation in Eq. (6) acts on all t token vectors, not on their mean. The second-update operation is therefore tokenwise across the complete update span, whereas the pre-answer operation has $t = 1$. We evaluate $\alpha \in \{0, .25, .5, .75, 1\}$ in Eq. (6); $\alpha = 1$ is full replacement, and the intermediate values trace the response between the unaltered target and full replacement.

Let $\widetilde{m}^{a\leftarrow b,r}_{i,M,Q,\ell}(\alpha)$ be the margin after intervening in target history H_a. Its change from the unaltered target is

$$\delta^{a\leftarrow b,r}_{i,M,Q,\ell}(\alpha) = \widetilde{m}^{a\leftarrow b,r}_{i,M,Q,\ell}(\alpha) - m^{r}_{i,M,Q}(H_a).$$

Set $s_{12\leftarrow 21} = +1$ and $s_{21\leftarrow 12} = -1$ so that both directions follow the $H_{21} - H_{12}$ orientation. The control-adjusted, direction-aligned margin change and matched-instance contrast are

$$\Gamma^{a\leftarrow b}_{i,M,Q,\ell}(\alpha) = s_{a\leftarrow b}\left(\delta^{a\leftarrow b,A}_{i,M,Q,\ell}(\alpha) - \delta^{a\leftarrow b,U}_{i,M,Q,\ell}(\alpha)\right), \qquad \Psi^{a\leftarrow b}_{i,\mathrm{int},\ell}(\alpha) = \mathcal{C}\big(\Gamma^{a\leftarrow b}_{i,\ell}(\alpha)\big).$$

The finite held-out-set effect is

$$\widehat{\theta}^{a\leftarrow b}_{\mathrm{int},\ell}(\alpha) = \frac{1}{24}\sum_{i=1}^{24} \Psi^{a\leftarrow b}_{i,\mathrm{int},\ell}(\alpha). \tag{16}$$

The normalized sample effect is

$$\widehat{\eta}^{a\leftarrow b}_{\mathrm{int},\ell}(\alpha) = \frac{\widehat{\theta}^{a\leftarrow b}_{\mathrm{int},\ell}(\alpha)}{\widehat{\theta}}. \tag{17}$$

This ratio divides the finite intervention effect by the observed H1 value for the same model. Its interval is obtained by dividing the endpoints of the intervention effect's matched-instance bootstrap interval by the fixed observed $\widehat{\theta}$. It therefore describes a scaled intervention effect conditional on the observed denominator and is not an interval for a jointly estimated ratio. On this scale, one means that the direction-aligned intervention interaction equals the observed H1 interaction in magnitude. It does not mean complete recovery of donor answers or explain a fraction of every individual margin.

The intervention tests influence on the specified behavioral contrast under residual-stream replacement, rather than mediation of the unmodified model's answer.

At full replacement, the pre-answer effect in each direction must have an unnormalized confidence interval strictly above zero and at least 18 positive matched-instance effects; each normalized sample effect must be at least .20. At the second-update span, both directional normalized intervals must lie strictly inside $[-.20, .20]$. The two directions and two locations are assessed with separate criteria; no direction difference or formal interaction between location and intervention effect is tested.

D.6 Selectivity criteria

Selectivity is evaluated for full replacement at the pre-answer position. Let $\nu \in \{12 \leftarrow 21, 21 \leftarrow 12\}$ denote the direction, $a(\nu)$ its target history, and c either the matched opposite-history intervention or a comparison condition. Write $\widetilde{m}^{\nu,c,r}_{i,M,Q}$ for the margin after applying c, and define the direction-aligned movement

$$\omega^{\nu,c,r}_{i,M,Q} = s_\nu\left\{\widetilde{m}^{\nu,c,r}_{i,M,Q} - m^{r}_{i,M,Q}(H_{a(\nu)})\right\}.$$

For question role r, define

$$C^{\nu,c,r}_i = \mathcal{C}(\omega^{\nu,c,r}_i), \qquad B^{\nu,c}_i = C^{\nu,c,A}_i - C^{\nu,c,U}_i, \qquad \widehat{\beta}^{\nu,c} = \frac{1}{24}\sum_{i=1}^{24} B^{\nu,c}_i.$$

For the matched opposite-history intervention, denoted $c = \mathrm{match}$, $\widehat{\beta}^{\nu,\mathrm{match}} = \widehat{\theta}^{\nu}_{\mathrm{int,ans}}(1)$.

Matched control questions. The collateral ratio asks how much matched-control margins move relative to update-relevant margins, irrespective of sign. The structured-control share instead asks how much the control questions contribute to the signed presentation–question-family interaction. Thus, broad absolute disturbance and disturbance with the interaction's structure are assessed separately. Define the cell-balanced absolute movements

$$A_i^\nu = \frac{1}{4}\sum_{M,Q} |\omega_{i,M,Q}^{\nu,\mathrm{match},A}|, \qquad U_i^\nu = \frac{1}{4}\sum_{M,Q} |\omega_{i,M,Q}^{\nu,\mathrm{match},U}|.$$

The collateral ratio is

$$R_{\mathrm{coll}}^\nu = \frac{24^{-1}\sum_i U_i^\nu}{\max(24^{-1}\sum_i A_i^\nu, 10^{-12})}.$$

Let $\overline{C}^{\nu,r} = 24^{-1}\sum_i C_i^{\nu,\mathrm{match},r}$. The structured-control share is

$$R_{\mathrm{struct}}^\nu = \frac{|\overline{C}^{\nu,U}|}{\max(|\overline{C}^{\nu,A}| + |\overline{C}^{\nu,U}|, 10^{-12})}.$$

Each quantity must be at most .33 in both directions. The first measures absolute movement on matched controls across the four cells; the second measures the control component of the signed interaction.

Identity check. Replacing a target representation with itself must leave the margin and every representation coordinate unchanged up to 10^{-6}. This checks the replacement operation rather than testing relation-specific influence.

Random perturbations and alternative donors. For five fixed random-direction comparisons, we add activation perturbations with the same Euclidean norm as the matched donor–target difference. Their fixed seeds are 731, 1459, 2903, 5801, and 11617. The same-answer comparison uses a donor from a different matched graph instance with the same correct answer and matched presentation, question family, and question role. The different-question comparison uses a different question from the same matched graph instance while retaining the presentation and question family.

For each of these seven comparison controls c and each direction,

$$R_c^\nu = \frac{|\widehat{\beta}^{\nu,c}|}{\widehat{\beta}^{\nu,\mathrm{match}}}, \qquad \widehat{\beta}^{\nu,\mathrm{match}} > 0.$$

Every ratio must be finite and at most .33; the five random-direction comparisons are assessed separately.

Cross-family comparison and availability. The cross-family comparison asks whether a history difference estimated from one question family can produce the corresponding effect in the other. It is a positive-transfer requirement, unlike the small-effect comparison controls above. For target instance i and family Q, let Q' be the other family. Suppressing the fixed model and pre-answer location, the prescribed reference is

$$D_{-i,M,Q',r} = \frac{1}{23}\sum_{j\neq i} \left[Z_{j,M,Q',r}(H_{21}) - Z_{j,M,Q',r}(H_{12})\right].$$

It averages within-source history differences equally across the remaining 23 instances while retaining the target presentation and question role. The operation adds $D_{-i,M,Q',r}$ to an H_{12} target or subtracts it from an H_{21} target. It does not replace the target with another family's state or rescale the reference to match a perturbation norm. Both source histories must match the target's decoder layer, tensor shape, selected-token count, and absolute token positions. The complete leave-one-instance-out source set is required for every target condition.

This control was prespecified only at the single-token pre-answer position. Under the exact signature above, no target condition has a complete 23-instance source pool, so every requested cross-family cell is recorded as unavailable rather than computed from a partial pool. If the required pool were available, the comparison would have to show a positive confidence interval, at least 18 positive matched-instance effects, a ratio to the observed H1 estimate of at least .20, a collateral ratio at most .33, and a structured-control share at most .33, separately in each direction.

A selectivity conclusion requires the pre-answer intervention criterion, both matched-control criteria, the identity criterion, all seven comparison-control ratios, the cross-family criterion, and observations for every required comparison. A comparison that cannot be formed does not satisfy its criterion.

The .20 and .33 bounds, the minimum of 18 positive instances out of 24, the 10^{-6} identity tolerance, and the use of five fixed random-seed directions are prespecified operational criteria. Endpoint and selectivity decisions apply to these bounds, internal locations, donor rules, and controls.

E Complete Behavioral Results

E.1 Confirmatory comparisons

The core study supports the prespecified update-order interaction in both models. Each graph instance first contributes a signed contrast. H1 averages these contributions; H2/H3 compare absolute values only after averaging the signed contributions within each description condition. Resampling repeats the complete operation, keeping all conditions from an instance together. Table 4 reports every held-out H1–H3 comparison. For H1, 21 of 24 instance contributions have the prespecified orientation in each model. The confidence intervals and multiplicity-adjusted tests, rather than these descriptive counts, determine the confirmatory decisions.

For H2, both additional final-state descriptions reduce the absolute aggregate contrast for Qwen3-VL. For Gemma, the correct-image comparison is supported and the correct-text comparison is inconclusive. H3 supports a difference between stale and correct images in both models. This last comparison remains composite because the two renderings were not matched feature by feature.

Table 4: Complete held-out behavioral comparisons. Estimates are $\widehat{\theta}$ for H1, $\widehat{\tau}_F$ for H2, and $\widehat{\xi}$ for H3. Intervals resample the 24 matched graph instances; p values are Holm adjusted within each hypothesis's prespecified comparison family.

Hypothesis	Model	Condition	Estimate	95% CI	Holm p	Decision
H1	Qwen3-VL	state-equivalent histories	5.781	[3.625, 8.073]	1.69×10^{-5}	supported
H1	Gemma	state-equivalent histories	3.671	[2.581, 4.855]	2.62×10^{-6}	supported
H2	Qwen3-VL	correct text description	9.258	[5.268, 13.790]	1.80×10^{-4}	supported
H2	Qwen3-VL	correct final-state image	11.455	[8.104, 14.892]	2.38×10^{-6}	supported
H2	Gemma	correct text description	0.787	[-0.891, 2.449]	0.188	not supported
H2	Gemma	correct final-state image	2.740	[0.711, 4.634]	0.0283	supported
H3	Qwen3-VL	stale vs. correct image	2.142	[0.557, 3.620]	0.00467	supported
H3	Gemma	stale vs. correct image	7.303	[4.153, 10.399]	3.42×10^{-4}	supported

H1 concerns a continuous correct-answer margin computed from teacher-forced complete answers. It does not require the model's selected answer to change. The core inference is conditional on the 24 held-out matched instances and the K_4 construction; its connection to final-state sufficiency is derived in Appendix D.1. Evidence on larger, non-complete endpoint graphs is reported separately in Appendix F.

E.2 Decomposition of the primary interaction

The same positive interaction arises from different condition-level patterns in the two models. Table 5 separates the update-relevant and matched-control order contrasts before showing their difference in each presentation–question condition. These components explain the aggregate; they are not separately tested hypotheses.

For Qwen3-VL, the degree-minus-shortest-path contrast is positive in text and negative in images, so subtracting the latter gives a positive interaction. For Gemma, the large negative text shortest-path component enters with a negative coefficient and makes the text family difference larger than the image difference. Both models thus support the fixed comparison without sharing a common sign pattern across conditions. This interpretation follows from the coefficients of $\mathcal{C}$, not from averaging the four components or requiring each one to be positive.

Table 5: Cell decomposition of the primary state-equivalent-history contrast $\widehat{\theta}$. For each initial-graph presentation and question family, Δ^A and Δ^U are the mean $H_{21} - H_{12}$ changes in the correct-answer margin for update-relevant and matched-control questions, respectively, and $E = \Delta^A - \Delta^U$.

Model	Initial-graph presentation	Question	Δ^A	Δ^U	E
Qwen3-VL	text	degree	0.688	-0.885	1.573
Qwen3-VL	text	shortest path	-1.865	0.302	-2.167
Qwen3-VL	image	degree	0.083	0.292	-0.208
Qwen3-VL	image	shortest path	0.771	-1.063	1.833
Gemma	text	degree	0.515	-0.095	0.609
Gemma	text	shortest path	-3.099	0.524	-3.623
Gemma	image	degree	0.172	0.022	0.150
Gemma	image	shortest path	-0.469	-0.057	-0.412

E.3 Signed aggregates in the final-state experiment

H2 and H3 compare absolute magnitudes only after the signed matched-instance contributions have been averaged. This differs from averaging the absolute reduction within each instance. Table 6 reports both summaries and names the reference and comparison conditions explicitly.

Table 6: Two distinct summaries of final-state-description effects. Arrows identify reference r and comparison c. The signed columns give $\widehat{\theta}_r^{\text{final}}$ and $\widehat{\theta}_c^{\text{final}}$; aggregate reduction is $|\widehat{\theta}_r^{\text{final}}| - |\widehat{\theta}_c^{\text{final}}|$, the primary H2/H3 estimand. Its inference is reported in Table 4. The last column instead gives the mean of $|\Psi_{i,r}^{\text{final}}| - |\Psi_{i,c}^{\text{final}}|$ and its descriptive 95% bootstrap interval. Text/image name the appended correct description, not the initial presentation; "stale" means a stale image.

Model	Reference → comparison	Signed reference	Signed comparison	Aggregate reduction	Instancewise reduction 95% CI
H2: no description versus a correct description					
Qwen3-VL	none → text	12.865	3.607	9.258	6.738 [2.532, 11.504]
Qwen3-VL	none → image	12.865	1.410	11.455	9.096 [5.975, 12.556]
Gemma	none → text	3.459	2.673	0.787	1.595 [0.334, 2.876]
Gemma	none → image	3.459	-0.719	2.740	2.097 [0.592, 3.688]
H3: stale versus correct image					
Qwen3-VL	stale → image	3.552	1.410	2.142	1.351 [-0.059, 2.825]
Gemma	stale → image	8.022	-0.719	7.303	6.028 [3.402, 8.732]

The distinction is clearest for Gemma with a correct final-state image: the signed aggregate changes from 3.459 without an additional description to $-.719$ with the image. Its absolute magnitude therefore falls to .719, even though the signed contrast crosses zero and individual contrasts need not move monotonically toward zero. The instancewise column is a different descriptive aggregation, not a robustness test of the primary estimand. This distinction also explains the Gemma correct-text result: its mean instancewise reduction has an interval above zero, while the prespecified aggregate attenuation is inconclusive. Only the aggregate estimand and its corresponding interval and paired test determine the H2/H3 decisions in Table 4.

This experiment uses the graph-changing-versus-state-preserving contrast, not the reversed-order contrast tested by H1. Attenuation therefore concerns a separate diagnostic rather than a retest of H1. The correct text and image conditions also differ in format and do not isolate a presentation effect.

For H3, the stale image depicts the graph-changing condition's pre-update state, whereas the correct image depicts its final state. Because the renderings also differ visually, H3 measures image-condition sensitivity rather than an isolated effect of graph-state content; Appendix B.3 describes the conditions.

F HELD-OUT TOPOLOGY TRANSFER BEYOND K_4

The controlled K_4 study isolates update order while holding the final graph fixed, but it cannot determine whether the resulting behavioral witness depends on that dense endpoint. We therefore evaluate the same necessary equality on an independently held-out set of larger, non-complete graphs. The extension preserves the matched update histories while varying graph family and size; its joint test is distinct from the directional interaction $\widehat{\theta}$ used for H1. This section describes the construction, calibration requirements, inference, and results.

F.1 DESIGN AND SCOPE

The extension contains five graph families chosen to vary the structural role of the two updates: component bridging, cycle closure, a symmetry-breaking chord, a sparse-bottleneck bypass, and hub attachment. Each family is instantiated at $n \in \{6, 8\}$. Within every family–size template, two genuinely absent edges are added in the update histories H_{12} and H_{21} in opposite orders. The two update histories therefore pass through different intermediate graphs and reach the same non-complete final graph. This retains the defining state-equivalence comparison while removing the complete-graph endpoint used in the core construction.

For an explicit specification, write the nodes as $v_0, \ldots, v_{n-1}$, let P_n be the path in that order, and let $C_n = P_n \cup \{(v_0, v_{n-1})\}$. In each construction below, G_0 is the initial graph, e_1, e_2 are the absent update edges, and the shared endpoint is $G_0 \cup \{e_1, e_2\}$.

Component bridging. Initially v_0 and v_1 are isolated, while $v_2, \ldots, v_{n-1}$ form a path. The updates add $e_1 = (v_0, v_1)$ and $e_2 = (v_1, v_2)$, yielding P_n. Each addition joins distinct connected components in either order.

Cycle closure. Initially v_0 is isolated and $v_1, \ldots, v_{n-1}$ form a path. The updates add $e_1 = (v_0, v_{n-1})$ and $e_2 = (v_0, v_1)$, yielding C_n. In either order, the first addition connects the isolated node and the second closes the cycle.

Symmetry-breaking chord. The initial graph is $C_n \setminus \{(v_0, v_1)\}$. Adding $e_1 = (v_0, v_{n/2})$ and $e_2 = (v_0, v_1)$ yields a cycle with a chord between opposite nodes.

Sparse-bottleneck bypass. The initial graph is P_n. The updates add $e_1 = (v_0, v_{n-1})$ and $e_2 = (v_1, v_{n-1})$, yielding a cycle with a short chord. The new connections provide alternatives to paths through the middle of the initial chain.

Hub attachment. The initial edges are (v_0, v_1), (v_1, v_2), (v_0, v_3), and (v_0, v_j) for $j \geq 4$. The updates add $e_1 = (v_0, v_2)$ and $e_2 = (v_2, v_3)$. In the endpoint, hub v_0 is adjacent to every other node, and the non-hub edges (v_1, v_2) and (v_2, v_3) remain. Only the first of the two named update edges is incident to the hub.

All ten endpoints are connected and non-complete; some initial graphs are disconnected. These constructions specify the tested structural scope rather than a random sample of graphs.

Each family–size combination has one canonical unlabeled topology, yielding 10 topology templates. A held-out realization applies a deterministic node relabeling and, for image inputs, a rotation or reflection together with layout jitter. There are 32 held-out realizations for each template and hence 320 matched graph pairs. These realizations test stability under node relabeling and changes in visual layout; they are not 320 independently sampled topologies. All conclusions are therefore conditional on the 10 prespecified templates and their held-out surface realizations.

The primary battery crosses text and image presentations with degree and shortest-path questions. For each of these four presentation–question conditions, an update-relevant question is paired with a matched-control question from the same question family. Two additional edge-adoption diagnostics check whether the requested edges are present in the final graph, but they do not enter the order-contrast estimand. Crossing conditions, update histories, question roles, and diagnostics produces 7,680 evaluated rows per checkpoint, or 23,040 rows across the three checkpoints.

Calibration, development, and held-out realizations were generated from disjoint deterministic identifiers. Calibration uses eight realizations per family–size combination to determine whether the model answers the direct questions and incorporates both updates reliably enough for an inter-

pretable order contrast. Development uses four additional realizations for integrity and execution checks only. The held-out statistics use the remaining 32 realizations, after calibration eligibility has been fixed.

F.2 Eligibility for order-contrast inference

A margin contrast is difficult to interpret when a checkpoint cannot solve the underlying graph query or apply the stated updates. We assess these abilities on calibration examples that are independent of the held-out update-history comparison. Eligibility is determined separately for each checkpoint–family–size panel from four prerequisites: parsed response rate, direct-task competence for both question roles in every presentation–question condition, recognition of each added edge in both presentations, and adoption of each update under both update histories and both presentation formats. None of these quantities compares H_{12} with H_{21}, so the held-out order contrast cannot determine eligibility.

The three accuracy prerequisites use the highest-likelihood valid complete answer, scored by the same teacher-forced protocol as the held-out margins. Direct-task competence tests the four task queries on an explicit final graph; direct-edge recognition tests whether each added edge is present in that graph; adoption tests those edges after each update history. The frozen thresholds are .70, .90, and .85, respectively. Each accuracy stratum contains eight calibration realizations, requiring at least $6/8$, $8/8$, and $7/8$ correct candidate selections. Taking the minimum over the relevant strata prevents strong performance on one presentation or query from compensating for weak performance on another.

Parse reliability is assessed separately from greedily generated responses. Each presentation has 80 responses per panel: eight realizations, each with six direct-final queries and two edge queries after each of the two histories. The .98 threshold therefore requires at least $79/80$ parseable responses in each presentation. All panels achieved a parse rate of one. Table 9 reports the minimum rates used for all 30 eligibility decisions.

We conduct the held-out order-contrast tests only in panels satisfying every prerequisite. Outside this set, variation in the margin cannot be distinguished from inadequate task execution under the calibration design. Such a panel has no formal order-contrast decision; this status is not evidence for a zero contrast or for equivalence.

F.3 Control-adjusted order contrasts

Let g, f, and n index checkpoint, graph family, and graph size, and let $i \in \{1, \ldots, 32\}$ index a held-out realization. A condition $s \in \mathcal{S}$ specifies one presentation and question-family combination, where $\mathcal{S} = \{(T, d), (T, \mathrm{sp}), (I, d), (I, \mathrm{sp})\}$. Using the correct-answer margin from Eq. (2), define the update-relevant and matched-control order contrasts

$$\Delta^{A}_{gfnis} = m^{A}_{gfnis}(H_{21}) - m^{A}_{gfnis}(H_{12}), \tag{18}$$

$$\Delta^{U}_{gfnis} = m^{U}_{gfnis}(H_{21}) - m^{U}_{gfnis}(H_{12}), \tag{19}$$

and their control-adjusted order contrast

$$e_{gfnis} = \Delta^{A}_{gfnis} - \Delta^{U}_{gfnis}. \tag{20}$$

Final-state sufficiency implies that both order contrasts vanish, and therefore implies $e_{gfnis} = 0$ for every condition and realization.

The primary panel statistic combines evidence across the four conditions without requiring a common direction:

$$Q_E(g, f, n) = \sum_{s \in \mathcal{S}} \frac{\left(\sum_i e_{gfnis}\right)^2}{\sum_i e^2_{gfnis}}. \tag{21}$$

The supporting update-relevant statistic is

$$Q_A(g, f, n) = \sum_{s \in \mathcal{S}} \frac{\left(\sum_i \Delta^{A}_{gfnis}\right)^2}{\sum_i (\Delta^{A}_{gfnis})^2}. \tag{22}$$

Each term compares the squared sum in one condition with its sum of squares; a condition with a zero denominator contributes zero. This self-normalization makes both statistics insensitive to a common nonzero rescaling within a condition. They are two-sided joint statistics, not effects measured in margin units, and conditions with opposite signs can both contribute to rejection. The control-adjusted statistic Q_E is the primary endpoint because it separates the update-relevant order contrast from background order sensitivity measured by the matched control; Q_A records the corresponding unadjusted update-relevant contrast.

p-values use one million Monte Carlo matched sign flips, with master seed 20260831. One random sign multiplies the complete four-condition vector for each graph realization; it is also shared across checkpoints when that realization is pooled. This retains within-realization dependence. Validity additionally requires joint sign symmetry of these vectors and independent signs across graph clusters under the null; matching alone does not supply that assumption. With $B = 10^6$ draws and T statistics at least as large as the observed statistic, the reported value is $(T + 1)/(B + 1)$, so even a zero exceedance count gives a positive Monte Carlo p-value.

Holm correction covers all 30 prespecified panels separately for Q_E and Q_A, at familywise level .05. For the model-level omnibus, each eligible panel retains its own four condition terms: the observed statistic is the sum of those panel statistics, not a contrast formed by first averaging different topologies. The sign-flip calculation uses the union of their graph clusters. Omnibus values are Holm corrected across the three checkpoints, separately for the two endpoints.

Condition-level means in Figure 5 are accompanied by pointwise 95% percentile intervals from 10,000 graph-cluster bootstrap samples, with master seed 20260830. Each draw resamples complete realizations within a panel, retaining all four conditions together. These intervals describe condition-level heterogeneity and receive no simultaneous multiplicity correction. Formal panel decisions use the joint Q_E test rather than whether an individual interval excludes zero.

F.4 Evidence across topology families

Gemma met the calibration prerequisites in seven panels, and in every panel the joint test rejected the null that all four control-adjusted order contrasts are zero after Holm correction over the complete 30-panel family. As shown in Table 7, the evidence covers component bridging, cycle closure, and the symmetry-breaking chord at both graph sizes, as well as hub attachment at $n = 6$. The largest Holm-adjusted p_E value is 8.88×10^{-4}. Thus the margin's dependence on update order is not tied to the K_4 endpoint for Gemma.

Table 7: Held-out tests for the seven calibration-eligible Gemma panels. Q_E is the primary joint statistic of the control-adjusted order contrasts, and Q_A is the supporting update-relevant statistic. Raw Monte Carlo p values use one million matched sign flips; adjusted values use Holm correction over all 30 prespecified panels separately for each statistic.

Family	n	Q_E	Raw p_E	Holm p_E	Q_A	Raw p_A	Holm p_A
Component bridging	6	48.52	2.00×10^{-6}	5.20×10^{-5}	52.12	2.00×10^{-6}	5.00×10^{-5}
Component bridging	8	46.04	1.00×10^{-6}	3.00×10^{-5}	51.00	1.00×10^{-6}	3.00×10^{-5}
Cycle closure	6	21.24	3.70×10^{-5}	8.88×10^{-4}	18.42	2.57×10^{-4}	6.17×10^{-3}
Cycle closure	8	43.75	1.00×10^{-6}	3.00×10^{-5}	39.75	1.00×10^{-6}	3.00×10^{-5}
Symmetry-breaking chord	6	53.05	1.00×10^{-6}	3.00×10^{-5}	56.69	1.00×10^{-6}	3.00×10^{-5}
Symmetry-breaking chord	8	54.65	1.00×10^{-6}	3.00×10^{-5}	51.01	1.00×10^{-6}	3.00×10^{-5}
Hub attachment	6	26.49	2.00×10^{-6}	5.20×10^{-5}	43.93	1.00×10^{-6}	3.00×10^{-5}

Pooling the 28 presentation–question conditions from the seven eligible panels gives Gemma $Q_E = 293.732$ over 224 held-out graph clusters. The corresponding Monte Carlo p-value is $p_{\mathrm{MC}} = 1.0\times10^{-6}$, the resolution of the one-million-draw calculation, and the Holm-adjusted p-value across the three checkpoints is 3.0×10^{-6}. The supporting update-relevant omnibus similarly rejects with $Q_A = 312.91$ and adjusted $p_A = 3.0\times10^{-6}$.

The condition-level contrasts clarify what the joint decisions establish. All seven image shortest-path intervals lie below zero, and five of seven text shortest-path intervals do so. Degree conditions are more heterogeneous: for example, the text-degree contrast is positive for component bridging at $n = 6$, whereas several image-degree contrasts are negative, and the hub-attachment image degree

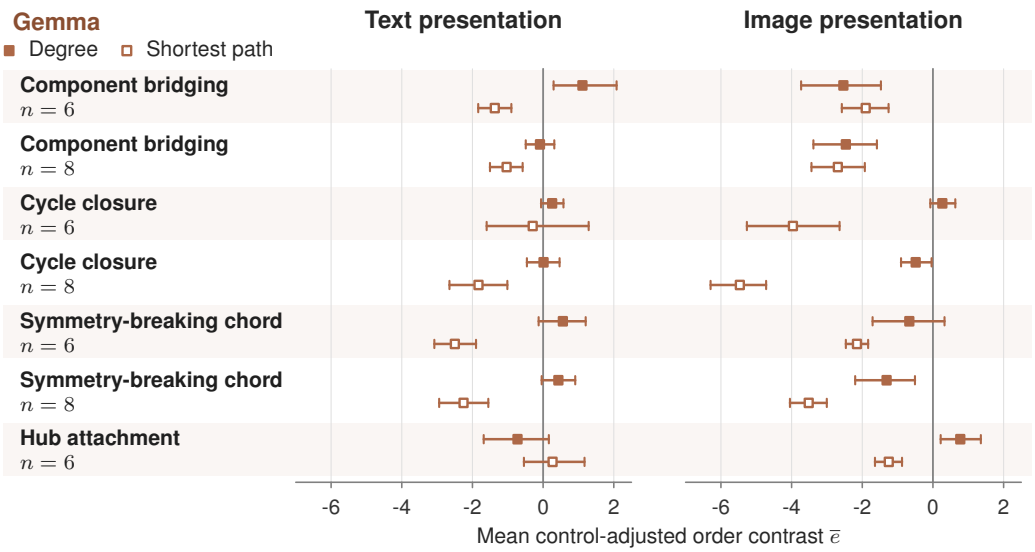

Figure 5: Condition-level contrasts in all seven eligible Gemma panels. Columns separate the initial text and image presentations on the same horizontal scale; filled and open squares denote degree and shortest-path questions. Points are means of $e = \Delta^A - \Delta^U$ across 32 held-out label/layout realizations, and bars are pointwise 95% graph-cluster bootstrap intervals. Both raw order contrasts use $H_{21} - H_{12}$: a negative e means that the relevant-question contrast is smaller than its matched control's contrast. Intervals are descriptive, without multiplicity adjustment; the formal panel decisions use the joint Q_E test in Table 7.

contrast is positive. The extension therefore supplies panel-level evidence across several topologies even though the presentation–question conditions need not share one direction.

F.5 Eligibility across checkpoints

Table 8: Held-out decisions and calibration eligibility for all 30 checkpoint–family–size panels. A green check marks a panel that met every calibration prerequisite and rejected the equality tested by the primary Q_E statistic after 30-panel Holm correction. Gray text names unmet prerequisites: C, direct-task competence; R, direct-edge recognition; A, update adoption. A gray-text entry has no formal order-contrast decision and is not a null result. CB, component bridging; CC, cycle closure; SC, symmetry-breaking chord; SB, sparse-bottleneck bypass; HA, hub attachment.

	Gemma		Qwen3.5		Qwen3-VL	
Family	$n = 6$	$n = 8$	$n = 6$	$n = 8$	$n = 6$	$n = 8$
CB	✓	✓	C/R	C/R	C	C/R/A
CC	✓	✓	C/R	C/R	C	C/R
SC	✓	✓	C	C/R	C	C
SB	C/R	C	C	C/R	C/A	C/A
HA	✓	R	C/R	C	C/A	C

Gemma is eligible in seven of ten panels; Qwen3.5 and Qwen3-VL each have zero eligible panels. Every Qwen panel falls below the direct-task competence threshold in at least one stratum. The unmet prerequisites are listed in Table 8, with underlying rates in Table 9. These panels do not provide a valid basis for distinguishing update-order dependence from insufficient downstream task competence.

For the multiplicity and breadth calculations, panels without eligibility receive $p = 1$ so that all prespecified panels remain in the analysis. This value is a fixed convention rather than a statistic computed from the held-out margin contrasts. These entries neither support equality nor show an absence of history dependence; the corresponding model–topology combinations remain unresolved.

Table 9: Calibration values underlying all 30 topology-transfer eligibility decisions. Each entry is the minimum over the prespecified strata, not a pooled accuracy. Parse uses 80 generated responses per presentation; task, edge recognition, and adoption use highest-likelihood valid-answer accuracy in eight-realization strata. Bold dark-gray values fall below the header threshold. A green check marks eligibility; a dash means no formal history decision, not a null effect. Bridge, Cycle, Chord, Bypass, and Hub abbreviate the five constructions in Appendix F.1.

Checkpoint	Family	n	Parse $\geq .98$	Task $\geq .70$	Edge $\geq .90$	Adoption $\geq .85$	Eligible
Gemma	Bridge	6	1.000	1.000	1.000	1.000	✓
Gemma	Bridge	8	1.000	1.000	1.000	1.000	✓
Gemma	Cycle	6	1.000	0.875	1.000	1.000	✓
Gemma	Cycle	8	1.000	0.875	1.000	1.000	✓
Gemma	Chord	6	1.000	0.875	1.000	1.000	✓
Gemma	Chord	8	1.000	0.875	1.000	1.000	✓
Gemma	Bypass	6	1.000	**0.000**	**0.875**	1.000	–
Gemma	Bypass	8	1.000	**0.625**	1.000	1.000	–
Gemma	Hub	6	1.000	0.750	1.000	0.875	✓
Gemma	Hub	8	1.000	0.750	**0.750**	1.000	–
Qwen3-VL	Bridge	6	1.000	**0.000**	1.000	0.875	–
Qwen3-VL	Bridge	8	1.000	**0.000**	**0.875**	**0.750**	–
Qwen3-VL	Cycle	6	1.000	**0.000**	1.000	0.875	–
Qwen3-VL	Cycle	8	1.000	**0.000**	**0.750**	0.875	–
Qwen3-VL	Chord	6	1.000	**0.000**	1.000	0.875	–
Qwen3-VL	Chord	8	1.000	**0.000**	1.000	0.875	–
Qwen3-VL	Bypass	6	1.000	**0.000**	1.000	**0.625**	–
Qwen3-VL	Bypass	8	1.000	**0.000**	1.000	**0.625**	–
Qwen3-VL	Hub	6	1.000	**0.125**	1.000	**0.750**	–
Qwen3-VL	Hub	8	1.000	**0.000**	1.000	0.875	–
Qwen3.5	Bridge	6	1.000	**0.375**	**0.875**	1.000	–
Qwen3.5	Bridge	8	1.000	**0.125**	**0.625**	0.875	–
Qwen3.5	Cycle	6	1.000	**0.250**	**0.875**	1.000	–
Qwen3.5	Cycle	8	1.000	**0.125**	**0.500**	1.000	–
Qwen3.5	Chord	6	1.000	**0.125**	1.000	1.000	–
Qwen3.5	Chord	8	1.000	**0.000**	**0.500**	1.000	–
Qwen3.5	Bypass	6	1.000	**0.000**	1.000	0.875	–
Qwen3.5	Bypass	8	1.000	**0.125**	**0.875**	1.000	–
Qwen3.5	Hub	6	1.000	**0.500**	**0.875**	0.875	–
Qwen3.5	Hub	8	1.000	**0.375**	1.000	1.000	–

F.6 Scope of the extension

The supported conclusion is a topology-scope result for Gemma. The joint null for the control-adjusted order contrasts is rejected in three non-K_4 families at both tested sizes and in a fourth family at $n = 6$. These held-out results rule out an explanation in which the Gemma witness is specific to the dense K_4 endpoint used in the core study. Together with the two-model core result, they extend the evidence along complementary model and topology dimensions.

We also specified a stronger nested breadth claim before held-out analysis: at least three of five graph families had to be supported in at least two of the three checkpoints, with both $n = 6$ and $n = 8$ supported for every contributing checkpoint–family pair. Let p_{gfn} be the raw Monte Carlo p-value for checkpoint g, graph family f, and node count n, with the fixed value one for ineligible panels. The nested calculation is

$$b_{gf} = \max\{p_{gf6}, p_{gf8}\}, \qquad c_f = \min\{1, 2b_{(2),f}\}, \qquad p_{\text{breadth}} = \min\{1, 3c_{(3)}\}.$$

The maximum requires support at both sizes. Here $b_{(2),f}$ is the second smallest of the three checkpoint values within family f, and $c_{(3)}$ is the third smallest of the five family values. The latter two steps are Bonferroni partial-conjunction calculations valid under arbitrary dependence. They use raw panel values, not the separately reported 30-panel Holm-adjusted values. Applied separately to Q_E and Q_A, this rule gives $p = 1.0$ for each. The breadth criterion is therefore not supported.

Because ineligible panels enter the nesting with the procedural value $p = 1$, this result limits cross-checkpoint breadth but does not show an absence of history dependence in those panels.

Inference is conditional on the ten fixed topology templates and their label and layout variants. The joint tests allow heterogeneous condition signs; their rejection does not imply the positive directional K_4 interaction. The extension measures complete-answer likelihoods, with free generation used only to check parsing. Its scope is behavioral; the internal interventions belong to the core study.

The archived specification also included an equivalence bound that was not used in the reported analysis; no equivalence claim is made.

G Relational Agreement Results

For each graph instance and initial presentation, we form four history-difference vectors per model: two question families crossed with update-relevant and control roles. Relational agreement is the Spearman correlation between the models' six pairwise cosine dissimilarities at a fixed location. Thus the analysis compares the organization of history differences without explicit alignment of hidden-state coordinates.

The prespecified comparison asks whether agreement is greater at the second-update span than at the pre-answer position. The paired differences are .905 (95% CI [.774, 1.043]) for text and .988 ([.821, 1.147]) for images. Both have Holm-adjusted $p = 1.19 \times 10^{-7}$. The pairing uses the same 24 matched graph instances at the two internal locations, and the bootstrap resamples those matched instances rather than individual conditions or pairwise distances.

Table 10: Complete relational-agreement results. Location estimates average cross-model Spearman correlations between each instance's six within-model RDM distances. Confirmatory inference applies to the paired second-update-minus-pre-answer difference; location-specific correlations are descriptive. Intervals resample the matched graph instances.

Initial-graph presentation	Quantity	Estimate	95% CI	Holm p	Decision
text	**second update − pre-answer**	**0.905**	**[0.774, 1.043]**	$\mathbf{1.19\times10^{-7}}$	**supported**
text	second-update correlation	0.964	[0.948, 0.979]	–	descriptive
text	pre-answer correlation	0.060	[-0.076, 0.188]	–	descriptive
image	**second update − pre-answer**	**0.988**	**[0.821, 1.147]**	$\mathbf{1.19\times10^{-7}}$	**supported**
image	second-update correlation	0.964	[0.949, 0.979]	–	descriptive
image	pre-answer correlation	-0.024	[-0.181, 0.138]	–	descriptive

The location-specific estimates provide the components of these differences. At the second-update span, cross-model relational agreement is .964 for text ([.948, .979]) and .964 for images ([.949, .979]). At the pre-answer position, it is .060 for text ([−.076, .188]) and −.024 for images ([−.181, .138]). These component estimates are descriptive: no location-specific test or equivalence-to-zero test was specified.

Agreement concerns the rank ordering of six dissimilarities, not identical features or causal use of those features. The supported location difference is conditional on the four tested conditions, frozen layers, and pooling rule specified in Appendix D.4.

H Intervention and Selectivity Results

H.1 Interpolation curves

The behavioral response to interpolation differs by location. At the second-update span, the direction-aligned interaction remains close to zero across the five strengths. At pre-answer, it grows as the matched donor state receives more weight, reaching .802–.960 of the observed H1 interaction at full replacement. Table 11 gives every point estimate. The directions exchange target and donor

histories within the same graph instance; their positive aggregates do not require each answer margin to move toward its donor value.

Table 11: Complete normalized intervention curves. Each cell is the direction-specific held-out estimate at the indicated α, divided by the same model's fixed observed H1 interaction. A value of one equals that interaction, not complete donor-answer recovery. Full-replacement endpoint intervals and prespecified decisions appear in Table 12. $12 \leftarrow 21$ denotes an H_{12} target with an H_{21} donor, and conversely for $21 \leftarrow 12$.

Model	Location	Direction	0	.25	.5	.75	1
Qwen3-VL	second-update span	$12 \leftarrow 21$	0.000	0.005	-0.005	0.002	-0.009
Qwen3-VL	second-update span	$21 \leftarrow 12$	0.000	-0.011	-0.004	-0.038	0.005
Qwen3-VL	pre-answer	$12 \leftarrow 21$	0.000	0.205	0.454	0.721	0.960
Qwen3-VL	pre-answer	$21 \leftarrow 12$	0.000	0.209	0.468	0.724	0.955
Gemma	second-update span	$12 \leftarrow 21$	0.000	0.009	0.001	-0.001	0.003
Gemma	second-update span	$21 \leftarrow 12$	0.000	-0.010	-0.019	0.001	-0.004
Gemma	pre-answer	$12 \leftarrow 21$	0.000	0.226	0.421	0.628	0.804
Gemma	pre-answer	$21 \leftarrow 12$	0.000	0.181	0.379	0.574	0.802

The intermediate strengths describe the response along the fixed interpolation path; no monotonicity test was specified. Confirmatory decisions use full replacement and the location-specific criteria in Appendix D.5.

H.2 Full-replacement endpoint decisions

Full replacement meets the prespecified location-specific criteria in both models (Table 12). At the second-update span, all four normalized intervals lie strictly within $[-.20, .20]$, satisfying the equivalence criterion in both directions for both models. This decision is limited to the tested residual-stream replacement and does not exclude other causal roles for the update representation.

Table 12: Full-replacement intervention endpoints and prespecified decisions. The second-update equivalence criterion requires both directional normalized 95% intervals for a model to lie strictly within $[-.20, .20]$. The pre-answer causal-influence criterion requires, in each direction, a positive raw-effect interval, at least 18 of 24 positive matched-instance effects, and a normalized point estimate of at least .20. Normalized intervals scale the raw bootstrap endpoints by the fixed observed H1 estimate; the denominator is not resampled jointly with the numerator.

Second-update equivalence

Model	Direction	Normalized estimate [95% CI]	Equivalence decision
Qwen3-VL	$12 \leftarrow 21$	-0.009 $[-0.041, 0.023]$	equivalent
Qwen3-VL	$21 \leftarrow 12$	0.005 $[-0.031, 0.040]$	equivalent
Gemma	$12 \leftarrow 21$	0.003 $[-0.024, 0.033]$	equivalent
Gemma	$21 \leftarrow 12$	-0.004 $[-0.029, 0.022]$	equivalent

Pre-answer causal influence

Model	Direction	Raw estimate [95% CI]	Positive /24	Normalized estimate [95% CI]	Causal-influence decision
Qwen3-VL	$12 \leftarrow 21$	5.552 $[3.573, 7.656]$	21/24	0.960 $[0.618, 1.324]$	supported
Qwen3-VL	$21 \leftarrow 12$	5.521 $[3.542, 7.594]$	21/24	0.955 $[0.613, 1.314]$	supported
Gemma	$12 \leftarrow 21$	2.952 $[2.312, 3.593]$	23/24	0.804 $[0.630, 0.979]$	supported
Gemma	$21 \leftarrow 12$	2.943 $[2.286, 3.638]$	24/24	0.802 $[0.623, 0.991]$	supported

At the pre-answer position, each direction has a positive unnormalized interval, at least 21 of 24 positive matched-instance effects, and a normalized point estimate above .20. Both models therefore satisfy the prespecified causal-influence criterion in both directions. Normalized intervals condition on the fixed observed H1 denominator, whereas the unnormalized contrast is the causal estimand (Appendix D.5). These are separate location-specific decisions, not a test of a location-by-intervention interaction.

H.3 SELECTIVITY CONTROLS

The available controls do not meet the joint selectivity criterion in either model. Table 13 separates three kinds of evidence: the identity check of the replacement operation; collateral and structured statistics of matched-control movement; and comparisons using random perturbations or alternative donors. Identity reinsertion has zero effect in every model–direction pair. Qwen3-VL remains below the .33 threshold for all ten random-direction comparisons, but exceeds it for the collateral and structured-control comparisons in both directions, both different-question donors, and both same-answer donors. Gemma exceeds the threshold for nine of ten random-direction controls, both different-question donors, one same-answer donor, and the collateral ratio in one direction. Thus, at least one available small-effect control exceeds its threshold in every model–direction pair.

Table 13: Complete pre-answer selectivity results. Identity entries report zero change, not a normalized ratio. Collateral compares absolute control movement with relevant-question movement; structured share measures the control contribution to the signed interaction. Random and alternative-donor entries divide their absolute aggregate effect by the matched-history effect in the same direction. Bold † marks a ratio above the prespecified .33 ceiling, not statistical significance. Random directions 0–4 use the five seeds listed in Appendix D.6, in that order. All 384 cross-family comparisons per model are unavailable; that comparison has a positive-transfer requirement rather than a small-effect ceiling. The left side of each arrow is the target history.

	Qwen3-VL		Gemma	
Check or statistic	12 ← 21	21 ← 12	12 ← 21	21 ← 12
Identity	0.000	0.000	0.000	0.000
Matched-control movement				
Collateral ratio	**1.015**†	**1.021**†	0.328	**0.343**†
Structured share	**0.460**†	**0.457**†	0.196	0.188
Random perturbations and alternative donors				
Random direction 0	0.092	0.045	**0.660**†	**1.796**†
Random direction 1	0.043	0.060	**0.439**†	**0.846**†
Random direction 2	0.026	0.008	**0.450**†	**1.829**†
Random direction 3	0.056	0.026	**0.986**†	**1.284**†
Random direction 4	0.011	0.053	0.063	**1.561**†
Same answer	**0.430**†	**2.777**†	0.119	**1.678**†
Different question	**1.927**†	**1.938**†	**1.542**†	**1.401**†
Positive transfer				
Cross-family	unavailable			

The planned cross-family control is unavailable under the frozen pre-answer design. Its additive reference vector requires exact agreement in layer, tensor shape, selected-token count, and absolute prompt positions. No complete 23-instance leave-one-out source pool from the other question family satisfies all four requirements, leaving 384 comparisons per model unavailable. An unavailable required control cannot provide evidence for selectivity. The negative selectivity decision does not depend on this missing comparison because multiple available controls already exceed the threshold for each model.

The ratios are point summaries over 24 held-out matched instances, not separately tested estimates. The joint rule, including the distinct positive cross-family requirement, is defined in Appendix D.6. The decision applies to the tested layer, donor construction, and control set.

I ADDITIONAL EXPERIMENTS ACROSS GRAPH FAMILIES AND SIZES

The earlier graph experiments compare a single graph-changing update with a state-preserving instruction, before and after a correct final-state description. They first screen direct graph-task competence, then examine where these update-condition contrasts recur. A follow-up varies the initial presentation and the final-state description to locate the conditions producing the contrast. Here, localization refers to these task conditions, not to an internal neural location. These exploratory studies motivate the later independent evaluations: the core study tests reversed update orders (H1), includes

a separate final-state-description diagnostic (H2/H3), and examines internal states; the topology extension tests reversed orders on larger graphs. The earlier observations do not enter any of these held-out estimates.

I.1 Relation among the study stages

The competence screen and exploratory comparisons are archived as Atlas Stages 1 and 2. An intermediate candidate-selection step, archived as Stage 3A, leads to the four-node dense-graph follow-up called Anchor-B. The later studies replace adaptive selection with frozen designs and independent held-out sets. Table 14 summarizes the studies reported below and their relation to the held-out evaluations. Archival coverage and verification notes are provided with the evidence files.

Table 14: Scientific roles of the graph studies. Discovery and localization guide study design but contribute no observations to the independent held-out estimates. Historical names identify the corresponding reports in the evidence archive.

Stage	Scientific role and scope	Evidence status
Atlas Stage 1	Screen history-independent competence across 26 family–size cells	Discovery screening
Atlas Stage 2	Explore update-condition contrasts in Stage-1-eligible cells	Adaptive discovery
Anchor-B localization	Vary initial presentation and final-state description on localization pairs	Follow-up on selected construction
Held-out K_4	Test H1; evaluate H2/H3 and internal diagnostics on independent instances	Frozen design; Section 4
Held-out topology extension	Test topology scope across five families and two sizes	Frozen design; Appendix F

I.2 Variable-size graph suite

The atlas crosses seven deterministic topology families with supported node counts in $\{4, 6, 8, 10\}$, producing 26 scenario–size cells. Each instance is a pair of graphs that differ by one target-edge change. Text and image presentations share the same graph semantics, while the query battery adapts to graph size and covers edge existence, degree, connectivity, shortest path, and component count. Table 15 states the intended structural role and admitted sizes of every family.

Table 15: Variable-size graph suite. Each scenario is a deterministic graph family; the listed node counts yield 26 fixed family–size combinations. The final column describes the structural role of the target-edge change.

Scenario family	Node counts	Structural role
Balanced bridge	4, 6, 8, 10	Component merge, reachability, and distance
Path–cycle closure	4, 6, 8, 10	Distance revision without a connectivity change
Triangle closure	4, 6, 8, 10	Local motif, endpoint degree, and distance
Hub attachment	4, 6, 8, 10	Shorter paths from a peripheral node through the hub
Symmetry-breaking chord	6, 8, 10	A chord changes cycle symmetry and distances
Bottleneck bypass	6, 8, 10	An additional connection between two dense subgraphs
Redundant dense edge	4, 6, 8, 10	Almost-clique completion with limited global change

The same edge change can merge components, create a local motif, alter route redundancy, or complete a dense subgraph. Results are therefore reported at scenario–size resolution rather than as a scaling trend. These 26 cells are distinct from the ten templates in the held-out extension, even where family names overlap; Appendix F.1 specifies that study's two-edge constructions.

I.3 HISTORY-INDEPENDENT COMPETENCE

The sealed atlas checkpoints are Qwen3.5-9B (Qwen Team, 2026) and Gemma-4-12B-it (Gemma Team, 2026). Both also appear in the held-out topology extension, which uses its own calibration and evaluation splits (Appendix F).

Stage 1 asks whether a checkpoint can recover a directly presented final graph before any condition contrast is examined. Eligibility is computed separately for every checkpoint–presentation–scenario–size cell from six graph-pair clusters and both edit directions. A cell must attain a parse rate of at least .98, edge-existence accuracy of at least .90, accuracy of at least .70 in every supported downstream query family, and complete required-role coverage. Each checkpoint report joins 21,120 direct-final prediction rows to the frozen source and records that no condition-contrast quantity was examined.

Coverage requires the prescribed questions to be present, not all answers to be correct. Each graph pair, presentation, and edit direction contains one target-edge query, $\binom{n}{2} - 1$ other adjacency queries, two update-relevant and two unchanged degree queries, one unchanged connectivity query, and one update-relevant and one unchanged shortest-path query. A further connectivity query and a component-count query are update-relevant when the edit changes components and unchanged otherwise. All roles must occur in both addition and removal conditions.

The six graph-pair replicates are partitioned in advance into indices 0–2 and 3–5. Stage-1 admission uses the full six-pair set; competence in each half is a stability diagnostic at that stage. Stage-2 candidate aggregation later requires both halves to pass competence. Before Stage 2, an additional class-balanced adjacency check requires both present-edge recall and absent-edge specificity to reach .90.

Table 16 discloses all 26×4 eligibility decisions. Qwen3.5-9B retains 7 image and 12 text cells; Gemma retains 10 image and 17 text cells. Seventeen cell–presentation combinations are eligible for both checkpoints. These counts define the available Stage-2 scope. They do not estimate a condition contrast, and an excluded cell contributes no such estimate.

I.4 EXPLORATORY CONDITION CONTRASTS

Stage 2 compares a graph-changing update with the same instruction applied to a graph in which the change is already satisfied. It evaluates both conditions without and with a complete, correct final-graph description. Within each graph pair and query family, the signed contrast is the state-preserving-minus-graph-changing correct-answer margin difference for update-relevant questions minus the corresponding difference for matched unchanged questions. Matched question groups are then averaged within the pair. Write this contrast as C^0 without a description and C^F with one. For graph-changing histories, a separate quantity S is the initial-state answer's log score minus the final-state answer's log score, averaged over update-relevant questions whose answers differ between states. It is not a correct-answer margin against the strongest alternative.

Five prespecified patterns summarize the exploratory results. A positive *pre-description contrast* concerns C^0, and a positive *post-description contrast* concerns C^F. A *signed contrast reduction* concerns $C^0 - C^F$ in a family already showing a positive C^0. Unlike H2, it does not test $|C^0| - |C^F|$: a signed decrease can cross zero.

Initial-answer preference requires positive S after a correct description in at least two eligible question families. *Cross-family preference disagreement* instead concerns opposite signs of S across families without a description. Neither preference pattern identifies a complete internal world state or requires the initial-state answer to be the highest-scoring answer among all candidates.

For traceability, the archived names of these five patterns are, respectively, "edit-compilation lag," "strict history," "refresh repair," "coherent refreshed stale candidate," and "query-family fragmentation." The behavioral names above describe the measured quantities rather than proposed mechanisms. The Qwen3.5 report contains 12,768 update-history free-generation rows and 11,400 likelihood rows; the Gemma report contains 18,144 and 16,200, respectively.

Prerequisites. Candidate support is evaluated separately by initial presentation and edit direction. An eligible downstream question family must have complete matched groups, parse rate at least .98, and candidate-argmax accuracy at least .70 under direct-final presentation and both state-preserving

Table 16: Complete Stage-1 competence matrix. Rows are the 26 fixed family–size cells and columns cross checkpoint with graph presentation. A green check means every direct-task criterion was met; a gray dash means at least one was not. The dash denotes ineligibility, not absence of an update-condition contrast. Eligibility uses six matched graph pairs and both edit directions without inspecting those contrasts.

		Qwen3.5-9B		Gemma-4-12B-it	
Scenario family	n	**Image**	**Text**	**Image**	**Text**
Balanced bridge	4	✓	✓	✓	✓
	6	–	–	✓	–
	8	–	–	–	–
	10	–	–	–	–
Path–cycle closure	4	–	✓	✓	✓
	6	–	–	–	–
	8	–	–	–	–
	10	–	–	–	–
Triangle closure	4	✓	✓	✓	✓
	6	✓	✓	✓	✓
	8	✓	✓	✓	✓
	10	✓	✓	✓	✓
Hub attachment	4	✓	✓	✓	✓
	6	✓	✓	✓	✓
	8	–	✓	–	–
	10	–	✓	–	–
Symmetry-breaking chord	6	–	–	–	✓
	8	–	–	–	✓
	10	–	–	–	–
Bottleneck bypass	6	–	–	–	✓
	8	–	–	–	✓
	10	–	–	–	✓
Redundant dense edge	4	–	✓	✓	✓
	6	–	✓	–	✓
	8	–	–	–	✓
	10	–	–	–	✓
Eligible cells		**7**	**12**	**10**	**17**

conditions. Edit adoption is checked on target-edge questions in the graph-changing conditions before and after the correct description: free-generation and candidate-argmax accuracy must each reach .85, and the median final-minus-initial answer log-score must be positive.

Positive contrasts and preferences. For the first four patterns, a question-family effect has stable positive support when all six pair contributions are present, at least four are positive, and the mean is positive in each fixed replicate half. In addition, its mean divided by its sample standard deviation must be at least .25 (or the variance must be zero with a positive mean). For pre- and post-description contrasts, a matched candidate-accuracy contrast of at least .10 is an alternative to this standardized-margin condition. This accuracy alternative is not used for signed contrast reduction or initial-answer preference. Signed reduction must satisfy its rule in the same question family as a supported pre-description contrast. Initial-answer preference requires stable positive support in at least two eligible families.

Opposite preferences across question families. Preference disagreement uses a different rule: in the graph-changing condition without a description, eligible families must have initial-minus-final score contrasts with opposite signs. Each included family must have an absolute standardized mean of at least .25, or a nonzero mean with zero variance. This rule does not require four positive pairs or positive means in both halves.

Promotion and counting units. For promotion to a later study, supporting entries must also pass competence in both fixed halves. At least four family–size cells from at least two graph families

must support a pattern within a fixed model, presentation, and edit direction. Promotion is adaptive nomination, not a confirmatory test of the later held-out estimands.

Table 17: Stage-2 exploratory patterns in the order defined in the text. Entries give supporting family–size cells, with the number of represented graph families in parentheses. Green bold entries meet the promotion rule; blue entries have support below that rule. A zero is an observed count, not a missing result. Columns separate initial presentation and edit direction. Qwen3.5 and Gemma denote the 9B and 12B-it checkpoints, respectively. These adaptive discovery counts are not independent confirmatory tests.

		Image		Text	
Checkpoint	**Exploratory pattern**	**Add**	**Remove**	**Add**	**Remove**
Qwen3.5	Pre-description contrast	2 (2)	0 (0)	**7 (3)**	2 (1)
	Post-description contrast	0 (0)	2 (2)	**6 (3)**	3 (2)
	Signed contrast reduction	2 (2)	0 (0)	**5 (2)**	2 (1)
	Initial-answer preference	0 (0)	0 (0)	0 (0)	0 (0)
	Cross-family preference disagreement	1 (1)	0 (0)	0 (0)	0 (0)
Gemma	Pre-description contrast	**4 (4)**	**5 (2)**	**13 (5)**	**13 (6)**
	Post-description contrast	1 (1)	**5 (3)**	**4 (2)**	**11 (5)**
	Signed contrast reduction	2 (2)	**4 (2)**	**13 (5)**	**10 (5)**
	Initial-answer preference	0 (0)	0 (0)	0 (0)	0 (0)
	Cross-family preference disagreement	0 (0)	0 (0)	0 (0)	0 (0)

Table 18: Report-level Stage-2 pattern flags. Counts include all flagged checkpoint–presentation–direction–family–size entries. Unlike the promotion counts in Table 17, they do not additionally require competence in both fixed replicate halves. Pre/post refer to the final-state description; reduction is signed, and the last two columns concern initial-versus-final answer preference.

Checkpoint	**Pre**	**Post**	**Reduction**	**Initial preference**	**Disagreement**
Qwen3.5-9B	16	14	13	0	1
Gemma-4-12B-it	38	24	31	0	0

Table 19: Cross-checkpoint support under text presentation with edge addition. Entries give supporting family–size cells, with graph families in parentheses. Shared counts their intersection, listed in the final column. This exploratory overlap motivates later confirmation rather than supplying it.

Pattern	**Qwen3.5**	**Gemma**	**Shared**	**Shared scenario–size cells**
Pre-description contrast	7 (3)	13 (5)	5 (3)	triangle n6,n8,n10; hub n6; redundant n4
Post-description contrast	6 (3)	4 (2)	4 (2)	triangle n6,n8,n10; redundant n4
Signed contrast reduction	5 (2)	13 (5)	4 (2)	triangle n6,n8,n10; hub n6

The report-level totals in Table 18 are 16 flagged pre-description contrasts, 14 post-description contrasts, 13 signed reductions, zero initial-answer-preference entries, and one preference disagreement for Qwen3.5; the corresponding Gemma totals are 38, 24, 31, zero, and zero. These totals count all flagged checkpoint–presentation–direction–family–size entries, including those that do not pass competence in both halves. By contrast, Table 17 counts supporting cells and graph families after the half-competence requirement, separately within each presentation and edit direction. Both checkpoints meet the promotion rule for pre-description contrast, post-description contrast, and signed reduction under text presentation with edge addition. Table 19 gives the cross-checkpoint overlap in that condition: five cells support the pre-description contrast in both reports, while four support each of the other two patterns. Neither report identifies a promotion-eligible initial-answer-preference pattern. This is a map of the discovery record, not evidence that the main K_4 interaction replicates across the atlas.

I.5 ANCHOR-B LOCALIZATION

The localization study varies initial presentation and final-state description for a four-node redundant-dense-edge addition. It uses Qwen3-VL-8B-Instruct (Bai et al., 2025) and Gemma before the confirmation study's held-out data were opened.

Preliminary competence. The accompanying checks assess graph-question accuracy, edge recognition, response parsing, and edge adoption. Each report covers eight graph pairs and 1,680 prompt conditions in total. Table 20 retains the reported rates. These are preliminary localization results, not measurements from the later confirmation study's development split. The archive does not recover per-metric denominators, complete row-selection rules, or the exact immutable model revisions for these earlier runs; 1,680 is therefore not the denominator of each rate.

Table 20: Archived preliminary competence checks from the localization study. All rates are percentages. Each report covers eight graph pairs and 1,680 conditions in total, not 1,680 observations for every metric; per-metric denominators are absent from the snapshot.

Model	Context-candidate accuracy	Direct edge	Response parsing	Target-edge adoption
Qwen3-VL	90.6	93.8	100.0	100.0
Gemma	99.5	100.0	100.0	100.0

Presentation and description contrasts. Let $s \in \{\text{text}, \text{image}\}$ denote the initial presentation and $t \in \{\text{none}, \text{text}, \text{image}\}$ the added final-state description. Here m_{noop} is the margin under a state-preserving instruction and m_{edit} under a graph-changing instruction. The shortest-path presentation–description contrast is

$$R_{s,t} = \big(m_{\text{noop}} - m_{\text{edit}}\big)_{\text{update-relevant}} - \big(m_{\text{noop}} - m_{\text{edit}}\big)_{\text{matched control}}, \tag{23}$$

formed within each graph-pair cluster before aggregation. It is the control-adjusted update-condition contrast without a description, or the remaining contrast after a correct description in the specified format. The contrast summaries use 16 matched pairs, distinct from the eight-pair competence summaries above.

Table 21: Shortest-path presentation–description follow-up (Anchor-B), using 16 matched graph pairs. Each entry reports the mean control-adjusted margin contrast, its 95% paired-bootstrap interval, and the number of positive pair contributions. The contrast is the state-preserving-minus-graph-changing difference for update-relevant questions minus that for matched controls. Bold intervals exclude zero descriptively, without a confirmatory multiplicity-adjusted decision. All other entries remain observed results, not unavailable comparisons. Text and image descriptions specify the correct final graph.

Initial presentation	Final-state description	Qwen3-VL	Gemma
Text	None	**-7.641 [-10.109, -5.187] 1/16 positive**	-1.609 [-4.223, 0.516] 7/16 positive
	Correct text	**-3.844 [-6.656, -1.109] 4/16 positive**	**-1.055 [-1.970, -0.149] 4/16 positive**
	Correct image	0.031 [-0.594, 0.625] 11/16 positive	**2.133 [1.293, 3.078] 15/16 positive**
Image	None	**4.453 [2.438, 6.828] 15/16 positive**	**3.200 [2.239, 4.121] 15/16 positive**
	Correct text	0.969 [-0.211, 2.281] 12/16 positive	**0.477 [0.223, 0.739] 13/16 positive**
	Correct image	**1.523 [0.406, 2.758] 13/16 positive**	-0.039 [-0.953, 0.891] 8/16 positive

Table 21 shows that the localization contrast varies across initial-presentation and final-state-description conditions, and its sign is not uniform across checkpoints. The table is therefore de-

scriptive evidence for choosing a later factorial comparison, not a universal estimate of history dependence. Only the two checkpoints with retained localization reports are included here; archival coverage is documented in the evidence README.

I.6 Relation to the held-out studies

Anchor-B and the main experiment share a four-node dense-graph context but use different contrasts. Atlas Stage 2 compares one graph-changing update with an idempotent state-preserving update. Anchor-B tests how initial presentation and final-state description affect that contrast. The main design instead starts from K_4 with two missing edges, applies both graph-changing additions in opposite orders, and compares update histories that reach the identical K_4. Its H1 estimand is the state-equivalent update-history interaction defined in Eq. (4), with the full specification in Appendix D.1; it is not an atlas signature. The held-out H2/H3 diagnostic retains a one-update comparison, but evaluates absolute aggregate attenuation and the stale-image contrast under its own frozen rules. It is not a confirmation of the atlas's signed-reduction flag.

The core study uses Qwen3-VL rather than the atlas's Qwen3.5-9B checkpoint. The topology extension preserves the reversed-order comparison on larger, non-complete templates. Both studies use independent held-out instances; neither pools observations from the atlas or Anchor-B.

J Evaluation Scope and Provenance

J.1 Evaluation scope

Core K_4 study. Per model, the 24 held-out matched instances supply 384 primary-history prompts, 576 no-description prompts in the separate one-update diagnostic, and 5,760 diagnostic prompts with descriptions. Each of the 384 activation records contains both internal locations. The 192 opposite-history prompt pairs yield $192 \times 2 \times 2 \times 5 = 3{,}840$ interventions: two directions, two locations, and five strengths. There are also 3,072 available control interventions and 384 unavailable cross-family comparisons per model. All repeat the same 24 analysis units.

Held-out topology-scope extension. Three checkpoints are evaluated on five graph families at $n \in \{6, 8\}$. The ten fixed templates each have 32 held-out label/layout variants, yielding 320 matched graph pairs and 7,680 evaluations per checkpoint, or 23,040 overall. Appendices F.1 and F.5 give the accounting by presentation, question, diagnostic role, and eligibility.

J.2 Development and held-out decisions

Core K_4 study. Separate development examples fixed the positive $H_{21} - H_{12}$ orientation. The two models, estimands, internal locations, intervention strengths, control rules, and statistical tests were fixed before the 24 held-out instances were evaluated once. The retained development record identifies the confirmation configuration. This was a private analysis plan, not a public preregistration. The earlier localization checks are reported separately in Appendix I.5.

Held-out topology-scope extension. The splits use distinct variant identifiers: eight calibration variants per family–size combination determine effect-blind eligibility, four development variants support integrity checks, and 32 held-out variants supply the final statistics. Families, sizes, presentation–question conditions, competence and edit-adoption prerequisites, joint statistics, multiplicity correction, and the cross-model breadth rule were frozen before held-out evaluation (Appendices F.1, F.3 and F.6).

J.3 Reproduction materials

The retained materials include model revisions, configurations, splits, eligibility decisions, and source analyses. File manifests and verification notes accompany the evidence files. Scientific specifications appear in Appendices B to D and F; Appendix I describes the separate exploratory records.