TRUE EVIDENCE, WRONG STATE: REPAIR OR ROLLBACK IN VISUAL COUNTERFACTUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual visual reasoning asks about an edited scene, while source images remain truthful about the scene before the edit. Rechecking this evidence can repair a wrong answer or roll a correct edited-state answer back to an obsolete one. In a local source-directed response, correction depends on how evidence offsets or repeats candidate error, captured by a cross-moment that permits opposite effects under identical marginal error distributions. Source strength controls whether and how long state risk improves, whereas query margins govern changes in the answer. Obsolete source displacement motivates state-compatible construction, while the interaction of construction quality and edit conflict motivates a separate source-retention decision. On MindEdit-L4, source evidence changes accuracy by -16.0/$+15.0$ percentage points for correct/unfinished post-state descriptions; fixed-bank exchanges isolate retention reversals, and separately trained replacement systems improve over append by 9.7–11.8 points across two families and two tasks. Evidence value depends on the queried state and candidate–source relationship, rather than on source truthfulness alone.

1 INTRODUCTION

Visual counterfactual reasoning asks what would hold after an imagined change to a scene. Revisiting an image can recover missed details (Bi et al., 2026; Ma et al., 2026), and imagined views provide intermediate evidence for spatial reasoning (Bigverdi et al., 2025; Yang et al., 2025). These gains concern the state represented by the evidence. An object-level edit changes the state that determines the answer (Sampat et al., 2021; Yu et al., 2026; Li et al., 2026). Moving a blue cube to the left of a red mug changes the leftmost object, although the source photograph remains accurate. When does truthful source evidence help an edited-state answer?

The same source can support repair and rollback. Ordinary verification on MindEdit-L4 corrects wrong answers but also returns correct answers to their pre-edit values. A fixed-description comparison exposes the dependence on candidate state (Fig. 1). For Qwen3-VL-4B, source-image evidence changes accuracy by -16.0 percentage points with correct post-state descriptions and $+15.0$ with unfinished-edit descriptions. Description quality is assessed independently of answer correctness; the source remains truthful in both conditions.

Visual revisiting retrieves observed details (Bi et al., 2026; Ma et al., 2026), while self-correction studies the reliability of feedback (Madaan et al., 2023; Huang et al., 2024; Gou et al., 2024; Wu et al., 2024). Perception-oriented training strengthens the evidence used for reasoning (Sharma et al., 2026; Venkatraman et al., 2026). Gains about an observed state do not automatically transfer to a changed target. Visual imagination and latent reconstruction create intermediate evidence (Bigverdi et al., 2025; Yang et al., 2025; 2026). These capabilities motivate holding construction fixed to isolate the value of original-source access.

Candidate error measures the need for correction, but does the source provide the missing information? We study candidate and source errors jointly, relative to the queried state. Their cross-moment measures error alignment, which determines the initial correction opportunity together with candidate risk. Changing this pairing can reverse verification value while preserving both marginal error distributions. Source strength can exhaust a benefit, while query margins determine answer changes. Raw movement toward the source and recovery followed by reapplication of the edit are distinct mechanisms; the latter can complete transformations that the former cannot.

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.

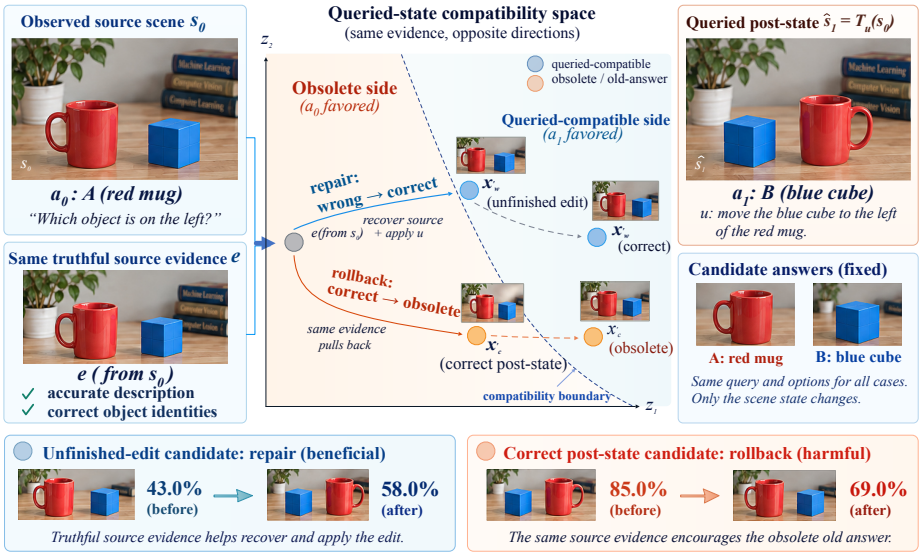


Figure 1: **The same truthful source can repair or undo an edit.** The query concerns the edited scene. Source evidence has opposite accuracy effects with correct and unfinished post-state descriptions. Photographs and trajectories illustrate state semantics; percentages are fixed-description accuracies, while native answer transitions are measured separately.

Source-state mismatch calls for evidence about the queried state, followed by a separate retention decision. An edit-conditioned constructor supplies banks for an empty-cache forward pass. With construction and reader fixed, reliable high-conflict states favor replacement, whereas incomplete states can benefit from append. Conflict changes this preference even at matched construction error. We test presentation and uncertainty explanations, then evaluate joint-error predictions on held-out scenes. Fixed-bank exchanges isolate retention, and separate system training tests the complete interface. Geometry, historical correspondence, and sequential edits then test state sufficiency.

2 When Does Source Verification Reverse?

Let S_0 be the observed scene, u the edit, and $S_1 = T_u(S_0)$ the queried scene. The question q defines the old answer $a_0 = q(S_0)$ and target answer $a_1 = q(S_1)$. We measure evidence value as its effect under a specified reader and verification interface. A wrong-to-correct answer transition is *repair*; a correct-to-a_0 transition is *rollback* when $a_0 \neq a_1$ and the old answer is identifiable.

Does rollback occur among ordinary requests? Native Qwen verification repairs 6.3% and damages 2.5%, for a net gain of 3.8 points, yet 1.3% of all requests return to an identified old answer. Positive average benefit thus conceals rollback (Appendix E.1). These paired transitions establish the event; fixed descriptions isolate its dependence on the supplied state.

The contrast persists when the source is explicitly labeled as pre-edit. After uncertainty matching, high conflict still produces a larger loss for correct post-state descriptions (Table 1). Randomizing the requested state to *before* instead makes the same source useful. Smaller non-state-contradiction and identity-two-pass effects distinguish state applicability from generic rechecking. The missing variable is how source information relates to the candidate and the requested world.

2.1 Changed States and Evidence Applicability

Hypothetical-action benchmarks already require reasoning about changed states (Sampat et al., 2021; Yi et al., 2020; Li et al., 2026). Structured scene representations and program execution provide ways to apply those changes (Yi et al., 2018; Mao et al., 2019; Gupta & Kembhavi, 2023; Surís et al., 2023; Reiter, 2001). The remaining question is conditional evidence value: given a candidate, when does its truthful source supply missing information or restore obsolete relations? We identify the error relationship that distinguishes these outcomes, then derive a state-compatible interface.

Table 1: **Source applicability under controlled comparisons.** Values are old-source minus neutral answer accuracy (pp). Conflict is matched on uncertainty; the target state is randomized in the next pair. Correct/poor always describes the post-state candidate.

Controlled comparison	Qwen3-VL-4B		Gemma-3-4B	
	Correct	**Poor**	**Correct**	**Poor**
Low conflict, matched	−2.0	14.0	−1.2	8.7
High conflict, matched	−15.1	13.2	−9.8	8.0
Randomized target: before	1.8	16.7	1.2	10.4
Randomized target: after	−15.7	15.0	−10.1	9.2
Non-state contradiction	−4.4	7.4	−2.5	4.9
Identity two-pass	−0.6	0.5	−0.4	0.3

3 What Determines Verification Value?

3.1 Candidate error is not repair opportunity

The controls point to state applicability. We isolate local source-directed movement to distinguish repeated from offsetting errors, then examine state-aware recomputation.

Write coordinates of the scene states S_0, S_1 as s_0, s_1. Let $s_1 = s_0 + d$, $X = s_1 + e$ be the candidate, and $O = s_0 + \nu = s_1 + z$ the source estimate, where $z = \nu - d$. Here e is candidate error, ν is source perception error, and z is source error *relative to the queried state*. Perfect perception removes ν but leaves the obsolete displacement $-d$. Consider

$$X_g = (1-g)X + gO, \qquad 0 \le g \le 1. \tag{1}$$

The scalar g measures geometric movement toward the raw source. For a fixed positive-semidefinite metric W and finite second moments, define

$$A = \mathbb{E}\|e\|_W^2, \qquad D = \mathbb{E}\|z\|_W^2, \qquad B = \mathbb{E}[e^\top W z]. \tag{2}$$

In Eq. (2), A measures candidate error and B measures error alignment, including covariance and mean bias. A source can therefore offset candidate error even when it is less accurate in isolation.

Proposition 1 (Repair opportunity and switching)**.** *For Eq. (1),*

$$R(g) = \mathbb{E}\|X_g - s_1\|_W^2 = (1-g)^2 A + 2g(1-g)B + g^2 D, \tag{3}$$

$$\Delta R(g) = R(g) - A = -2g(A-B) + g^2 H, \quad H = A + D - 2B \ge 0. \tag{4}$$

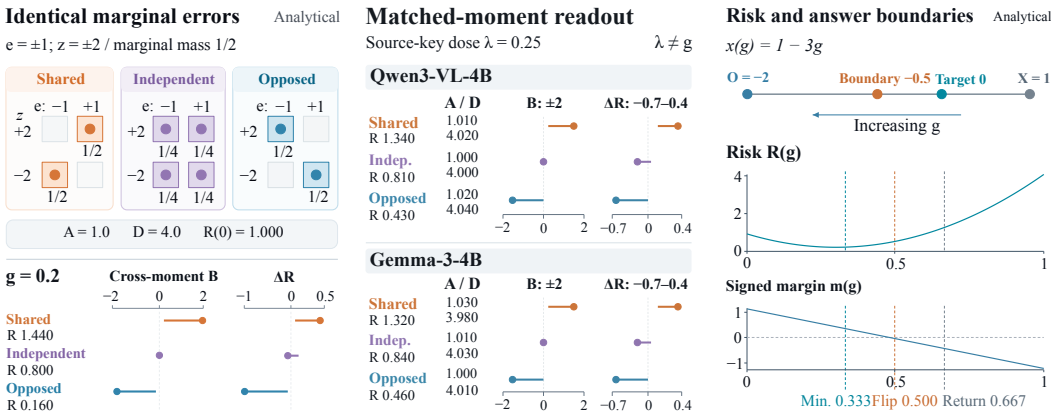
For $H > 0$, positive strength improves risk exactly when $A > B$ and

$$0 < g < g_{\text{switch}} = \frac{2(A-B)}{H}, \qquad g \le 1. \tag{5}$$

For $H = 0$, every strength has risk A.

Expanding $(1-g)e + gz$ gives Eq. (3). Correction opportunity is $A - B$, so alignment with source error reduces the initial benefit. The quadratic cost $H = \mathbb{E}\|e - z\|_W^2$ controls how far the response can exploit that opportunity. For a coordinate-correct candidate, $A = B = 0$ and raw source use adds $g^2 D$. For $H > 0$, the risk minimum is at $\text{clip}_{[0,1]}((A-B)/H)$, halfway to the switching point before clipping. Appendix A.2 gives the proof and degenerate cases.

Source truthfulness and perceptual accuracy therefore answer different questions from target relevance. With perfect perception, $\nu = 0$, source error is still $z = -d$. A coordinate-correct state has nothing to gain from movement toward that obsolete state; an incorrect candidate may either oppose or share its displacement. Even independent errors can have $B \neq 0$ when their means align. Neither candidate error magnitude nor source perception quality alone determines the initial correction. Next we hold both complete marginal error distributions fixed and vary only their pairing.



(a) Fixed marginals (b) Neural responses (c) Risk versus answer

Figure 2: **Error pairing reverses evidence value at fixed marginal quality.** (a) Analytical couplings preserve both marginal distributions. (b) Neural readouts approximately match selected moments at source-key dose $\lambda = 0.25$. (c) In an analytical query, the answer flips before risk returns to baseline. Attention dose λ differs from geometric strength g.

3.2 Identical Marginals, Opposite Verification Value

Can the two marginal error distributions determine whether the source helps? Figure 2(a) fixes $e \in \{-1, 1\}$ and $z \in \{-2, 2\}$ with equal probabilities, so $A = 1$ and $D = 4$ in every coupling. Pairing errors in the same direction gives $B = 2$ and $R(0.2) = 1.44$; independent pairing gives $B = 0$ and risk 0.80; opposed pairing gives $B = -2$ and risk 0.16. Both marginal distributions remain identical, yet pairing reverses the effect. Separate quality estimates cannot distinguish these responses. When do marginal risks suffice, and when is dependence necessary? The next bound identifies which conclusions are possible from marginal risks alone.

Proposition 2 (Sharp dependence region)**.** *For fixed A, D and $g \in [0, 1]$,*

$$\left((1-g)\sqrt{A} - g\sqrt{D}\right)^2 \le R(g) \le \left((1-g)\sqrt{A} + g\sqrt{D}\right)^2. \tag{6}$$

If $A > D$, every positive strength improves risk for every admissible coupling. If $D > A > 0$, both effect signs are attainable for $0 < g < 2\sqrt{A}/(\sqrt{A} + \sqrt{D})$; larger strengths harm every coupling.

Cauchy–Schwarz gives $|B| \le \sqrt{AD}$, with equality for aligned or opposed errors. At the threshold the best coupling ties the baseline; when $A = D$, no coupling increases risk within the mixture class. Dependence contributes $2g(1-g)B$ and disappears at $g = 1$, where raw-source risk is D. Figure 2(b) tests this ordering with approximately matched neural moments, while panel (a) fixes complete distributions. Section 5.2 specifies the neural intervention.

3.3 From State Risk to Answer Changes

An improvement in coordinate risk need not repair an answer. A query depends on which side of a decision boundary the state occupies, rather than distance alone. For a binary affine query, let the candidate have correct signed margin $c > 0$ and the old estimate margin $-a < 0$. Equation (1) gives

$$m(g) = (1-g)c - ga, \qquad g_{\text{answer}} = \frac{c}{a+c}. \tag{7}$$

In Fig. 2(c), $x_g = 1 - 3g$ has target zero and query threshold $-1/2$. Risk is minimized at $1/3$, the answer flips at $1/2$, and risk returns to baseline at $2/3$. Answer prediction therefore requires joint query margins beyond A, B, D. Multiclass errors need not restore the old answer.

Can raw-source movement complete an unfinished edit? For $X = s_1 - \alpha d$, $O = s_0$, and $0 < \alpha < 1$, raw mixing leaves error $-[\alpha + g(1-\alpha)]d$ and worsens every nonzero collinear under-edit. Recovering the source and reapplying the edit supplies the missing direction (Appendix B.2). This distinction motivates transporting source information before verification.

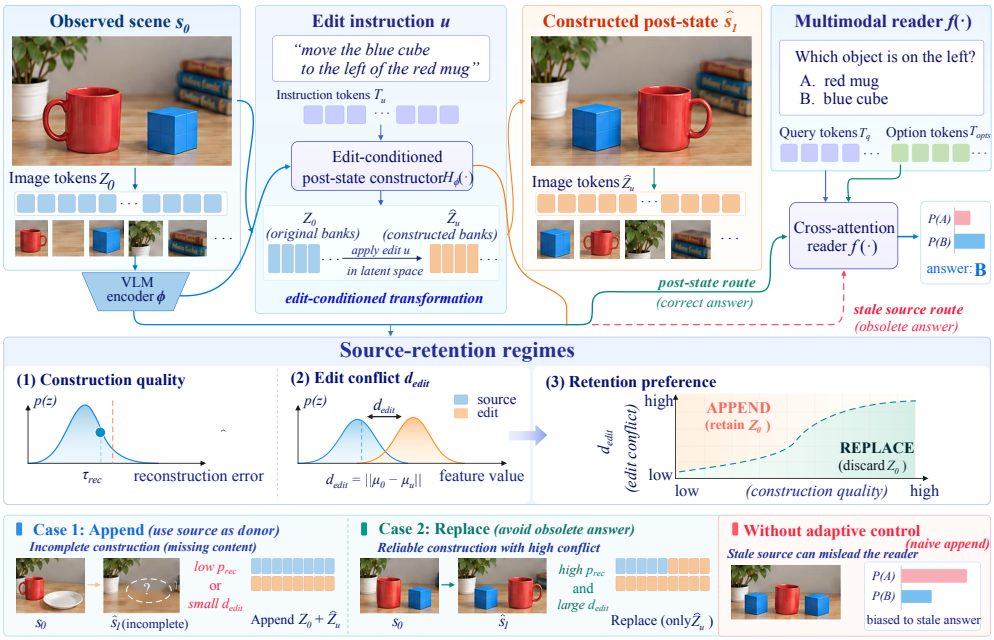


Figure 3: **State-compatible construction and source retention are separate decisions.** The constructor uses the edit before terminal reading. Append retains both bank sets; replacement exposes constructed banks alone. Scenes illustrate state semantics; lower panels summarize the quality–conflict regimes tested in Fig. 4. Appendix H.3 evaluates the separate public-input policy.

4 State-Compatible Evidence

4.1 Transport Evidence to the Queried State

Raw-source error $z = \nu - d$ contains obsolete displacement. Applying the edit before evaluating q, or evaluating $q \circ T_u$ on the source, aligns its state semantics (Reiter, 2001). For transported evidence $O_T = s_1 + z_T$, let $D_T = \mathbb{E}\|z_T\|_W^2$ and $B_T = \mathbb{E}[e^\top W z_T]$.

Proposition 3 (State transport)**.** *For fixed candidate and strength, with finite transported moments,*

$$R_{\text{raw}}(g) - R_T(g) = g^2(D - D_T) + 2g(1-g)(B - B_T). \tag{8}$$

For a fixed additive edit and independent zero-mean e, ν, the difference is $g^2\|d\|_W^2$.

Transport removes obsolete displacement in the additive case. Learned construction also changes error and dependence through D_T, B_T. The principle permits text, geometry, latent banks, or composed queries; we implement it with native visual banks (Fig. 3).

4.2 Construct the Post-State Before Reading It

Scene banks are the visual tokens injected into the language-model reader. A frozen text parser separates edit u from the terminal query. The constructor receives original banks Z_0 and edit encoding U_u, but neither the terminal query nor the answer options:

$$\widehat{Z}_u^k = Z_0^k + W_{\text{out},k} H_\phi(Z_0, U_u). \tag{9}$$

Four width-512 attention blocks implement H_ϕ with zero-initialized output projections (Vaswani et al., 2017). Terminal reading clears caches and preserves public options. Multiple queries reuse a construction; subsequent edits use predicted intermediate banks.

Frozen before/after encodings and answers supervise training:

$$\mathcal{L} = \tfrac{1}{2}\mathcal{L}_{\text{cf}} + \tfrac{1}{4}\mathcal{L}_{\text{post}} + \tfrac{1}{4}\mathcal{L}_{\text{obs}} + \beta\mathcal{L}_{\text{state}}. \tag{10}$$

The answer terms use predicted banks, training post-state banks, and original-scene factual queries, respectively; $\beta = 1$ weights state loss. Removing factual supervision lowers invariant accuracy from 86.3% to 83.7% (Appendix F.6). Appendix C specifies optimization and the native interface.

4.3 SEPARATE RETENTION FROM CONSTRUCTION

Construction determines the represented state; retention determines whether its original source remains accessible. Labeled append exposes both original and constructed banks, whereas replacement exposes the constructed banks alone. A fixed-bank exchange holds construction, reader weights, query, options, and empty caches constant. This exchange isolates source availability independently of learned construction and reading.

Figure 3 summarizes the quality–conflict regimes tested by this exchange. The main complete systems train fixed append or replacement interfaces; Appendix H.3 separately tests a public-input retention policy. Complete replacement changes every scene-related bank, including Qwen3-VL's secondary injections (Bai et al., 2025). It supplies a constructed post-state, unlike raw-source adoption at geometric $g = 1$. Unsupported requests retain original-image fallback through the same trained reader. The controls in Appendix F.3 test residual source access through caches and secondary banks.

5 EXPERIMENTS

5.1 SETUP

MindEdit-L4 evaluates object-level edits in natural photographs; STARE-rigid-text contains the public text-instructed 3D translation and rotation tasks (Yu et al., 2026; Li et al., 2026). Qwen3-VL-4B and Gemma-3-4B are the principal backbones (Bai et al., 2025; Gemma Team, 2025). Paired scenes supply training post-states and controlled coordinate readouts. Pairing and prediction test controlled responses; fixed-bank exchanges isolate retention; trained systems include native fallback. Appendix D specifies source-scene splits and comparison permissions.

5.2 ERROR PAIRING AND SOURCE STRENGTH

Does pairing change neural response when the two estimates have nearly the same marginal quality? Within matched canonical scene/edit structure, candidate replicas and source-readout errors are paired to form shared, independent, and opposed conditions. Selected marginal moments remain approximately matched (Fig. 2(b)). At source-key attention dose $\lambda = 0.25$, shared errors increase Qwen's coordinate risk by 0.33, while independent and opposed errors reduce it by 0.19 and 0.59. Candidate risk stays near one and source risk near four. Gemma reproduces this ordering, linking the response direction to which errors the source repeats or offsets.

Increasing source-key attention tests whether a useful correction remains useful at greater dose. The six-dose responses show independent errors first improving and later losing the benefit, while shared errors worsen from low dose (Appendix E.7). Attention dose λ is an intervention on source keys, not the geometric coefficient g; a development-fitted map relates them. Transport and recomputation separately improve collinear under-edits that raw movement worsens. These coordinate contrasts identify response mechanisms. Predicting actual answers additionally requires the query margins.

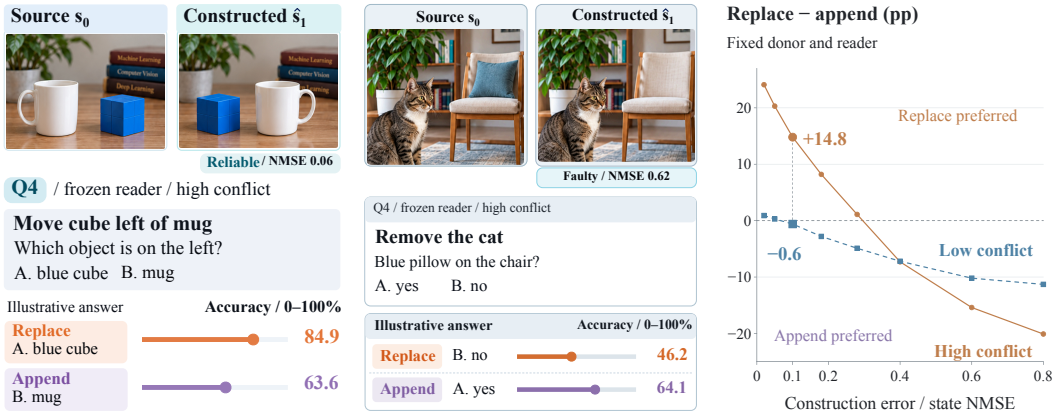
5.3 JOINT ERRORS IN HELD-OUT PREDICTION

Does joint dependence improve held-out answer prediction? All predictors receive matched controlled geometric inputs. We fit the dose–response map on development coordinates, propagate joint residuals through query margins, and calibrate correctness probabilities on source-disjoint development scenes. Independence controls preserve marginal distributions and means, either holding the other components fixed or refitting them. Flexible marginal-only and joint predictors test whether the result depends on the structural response map. Answer-correctness Brier loss includes every held-out case, including responses with no crossing (Table 2).

Refitting does not recover the information lost by removing dependence. On Qwen ID, independence refitting lowers Brier loss from 0.210 to 0.198, whereas the structural joint predictor reaches 0.184. The joint predictor also leads on asset shift. Flexible joint predictors instead lead on unseen composition and zero-calibration Gemma; family-specific calibration restores the structural model's advantage. Dependence remains useful when the structural response map requires recalibration (Appendix E.5). This prediction study uses shared geometric information. We next fix constructed banks and reader weights to test retention at the native interface.

Table 2: **Answer-correctness Brier loss ↓ on all held-out cases with matched geometric inputs.** Structural controls remove dependence; flexible predictors test response-map transfer. Bold/underline mark the two lowest point estimates per population.

Population	Structural response			Flexible predictors		
	Joint	**Indep. fixed**	**Indep. refit**	**Marginal**	**Joint GBDT**	**Joint MLP**
Qwen: ID	**0.184**	0.210	0.198	0.200	0.194	<u>0.192</u>
Qwen: asset shift	**0.195**	0.223	0.211	0.212	0.204	<u>0.202</u>
Qwen: composition	0.227	0.253	0.235	0.228	**0.214**	<u>0.216</u>
Gemma: zero calibration	0.219	0.247	0.236	0.222	**0.209**	<u>0.211</u>
Gemma: dev. calibration	**0.191**	0.218	0.204	0.207	0.201	<u>0.199</u>



(a) Reliable state (b) Poor state (c) Quality and conflict

Figure 4: **Quality alone does not determine the append–replace preference.** (a,b) With construction and reader fixed, reliable high-conflict states favor replacement and poor states favor append. Images illustrate the conditions; percentages are aggregate accuracies. (c) At matched construction error, edit conflict changes the retention benefit.

5.4 When should the original source remain?

The same construction can behave differently depending on source access. With banks and reader fixed, reliable high-conflict Qwen states favor replacement, 84.9% versus 63.6%, while poor states favor append, 64.1% versus 46.2% (Fig. 4(a,b)). At construction NMSE 0.10, replacement gains 14.8 points under high conflict but loses 0.6 under low conflict (panel (c)). Construction quality alone therefore cannot determine retention value. Original evidence may supply missing content or compete with already reliable edited relations.

Could a reader trained for one interface produce the reversal? Mode-balanced training improves append on reliable states but preserves its high-conflict deficit. Oracle donor banks remove construction error without eliminating reading errors, and padding-matched replacement stays close to ordinary replacement (Appendices F.1–F.2). Complete-bank, cache, and textual-history controls separately test stale information routes (Appendix F.3). These controls separate retention from reader training and residual source access. Complete-task evaluation next asks whether the interface remains useful when construction and reading are learned together.

5.5 Complete-task capability

Table 3 evaluates separately trained systems, including automatic parsing and native fallback. Relative to append, replacement gains 11.8/10.2 points for Qwen and 10.3/9.7 for Gemma on L4/STARE-rigid-text. This comparison tests the complete interface rather than holding the learned construction and reader fixed. Paired SFT, answer-only capacity, and objective controls separate the role of the interface from added parameters and supervision (Appendices F.4 and G.1).

Table 3: **Complete-task accuracy (%) ↑, including fallback.** L4 is MindEdit-L4; STARE denotes STARE-rigid-text. P uses paired images and answers; G adds geometric labels, F filtering experts, and E a geometry expert. Venue identifies the original method; all task-port scores use our protocol. Bold/underline rank only P; extra-resource rows are unranked.

Method / task port	Venue	Info.	Qwen3-VL-4B		Gemma-3-4B	
			L4	**STARE**	**L4**	**STARE**
Paired-data controls and task ports						
Paired SFT	–	P	33.1	51.7	28.7	47.0
Labeled append	–	P	46.0	62.0	39.9	56.8
Outcome GRPO	–	P	52.2	66.3	45.8	61.5
Mirage port	CVPR'26	P	<u>54.9</u>	<u>70.6</u>	<u>47.7</u>	<u>63.9</u>
CVPD port	BMVC'26	P	50.1	66.5	44.5	62.2
Predicted text	–	P	52.6	68.9	46.3	63.2
Post-state replacement (ours)	–	P	**57.8**	**72.2**	**50.2**	**66.5**
Additional labels or experts						
PVST task port	ECCV'26	P+F	49.8	66.8	44.2	61.6
Text with geometry labels	–	P+G	57.0	75.4	51.0	70.0
Predicted-geometry executor	–	E	41.7	70.2	38.9	67.7

The external task ports test related evidence strategies under the same evaluation protocol. Mirage constructs latent visual tokens, CVPD distills perception, and PVST verifies captions before regenerating reasoning (Yang et al., 2026; Venkatraman et al., 2026; Sharma et al., 2026). Replacement exceeds Mirage, the closest paired-data port, by 2.9/1.6 points on Qwen and 2.5/2.6 on Gemma. PVST uses filtering experts and belongs to the additional-resource block. Geometry-supervised text reaches 75.4% for Qwen and 70.0% for Gemma on STARE, above paired-data replacement. The evidence interface and additional spatial labels therefore provide different sources of improvement.

Within P, every method receives paired before/after images and answers during training, while inference retains the original public inputs. Paired SFT already uses the post-state targets without a constructor. Its comparison with replacement therefore tests how shared supervision is used. Filtering experts and geometric labels instead change the supervision budget, as marked separately in the table.

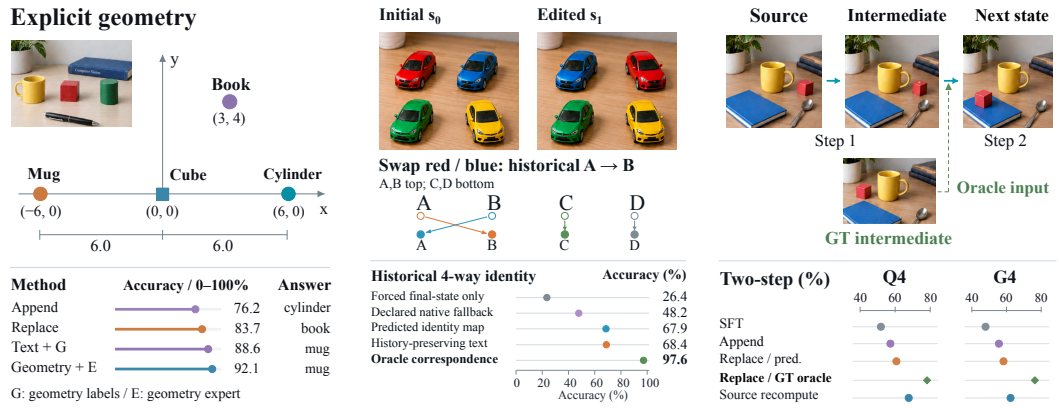
A predicted post-state description provides a simpler state-compatible alternative. Paired-data text improves over append in all four columns, but remains below latent replacement. For Qwen, its L4/STARE accuracies are 52.6/68.9 compared with 57.8/72.2 for replacement. Both methods construct evidence for the edited state, but their representations retain different information for the reader. The additional-label results show why representation must be compared at fixed supervision: a text representation can be stronger when it is trained with explicit geometry.

The paired-data comparison extends to Qwen3-VL-8B and InternVL3.5 (Wang et al., 2025b). Qwen-8B replacement reaches 65.8/76.6, against append's 55.0/67.1. On an independent real-photo cohort, replacement reaches 57.1%, append 47.4%, and geometry-supervised text 59.2%. Appendix G expands the family, task-port, and resource comparisons. A useful compatible state need not use a particular representation; the remaining question is which information that representation preserves.

5.6 What information must a useful state preserve?

Compatibility specifies the represented world; sufficiency specifies what can be answered from it. Figure 5 separates three requirements. Explicit geometry benefits rigid spatial queries. Historical references require identity correspondence: forced final-state-only reading gives 26.4% accuracy, versus 67.9% with a predicted identity map. The map preserves which object occupied an earlier role without treating its old location as current. Accurate final geometry alone cannot answer a question about which of two visually identical objects moved. State-indexed correspondence preserves that referent while allowing its current position to change (Appendix I).

Sequential edits need information beyond the first answer. On unseen compositions, replacement exceeds append by 2.7 points on Qwen and 0.5 on Gemma.



(a) Explicit geometry (b) Historical identity (c) Unseen composition

Figure 5: **State compatibility does not guarantee query sufficiency.** Queries may require (a) explicit geometry, (b) historical correspondence, or (c) information for subsequent edits. Scenes illustrate the queries; percentages summarize distinct populations and information budgets.

Source recomputation raises Qwen accuracy from 60.7% to 68.9%; an oracle intermediate reaches 74.2%. Reapplying the edit recovers information needed for the next transformation (Appendix H.2).

Table 4: **Accuracy (%) ↑ across information regimes.** The donor block fixes construction and reader; bold compares its retention modes. Remaining rows are unranked task-specific evaluations. P+G adds geometric labels; E uses a geometry expert.

Regime	Paired-data interface		Additional information	
	Append	**Replace**	**Text (P+G)**	**Geom. (E)**
Fixed-bank donor comparison				
Poor state, high conflict	**64.1**	46.2	64.8	67.9
Good state, high conflict	63.6	**84.9**	84.1	72.0
Distinct task and control populations				
Explicit rigid geometry	76.2	83.7	88.6	92.1
Historical referent (fallback)	48.8	48.2	68.4	65.3
Identical-input binary pairs	49.7	49.6	49.8	49.5
No edit	91.2	91.4	91.8	92.1
Already compatible	86.2	86.0	87.0	88.1
Strong spatial base	83.5	83.9	84.2	85.7
Unseen two-step edit	58.0	60.7	68.9	70.6

Table 4 distinguishes missing information from absent conflict. No-edit and compatible states leave little obsolete displacement to remove, whereas identical-input worlds conceal the target answer for every representation of the permitted inputs (Lemma B.4). A representation may discard information present in the source, or the permitted inputs may never reveal it. Preserving a historical referent addresses the former; changing the retention rule cannot resolve the latter.

6 Conclusion

Truthful source evidence can correct a counterfactual answer or restore an obsolete one. Joint error determines the correction available to a source-directed response; strength and query margins determine its consequences. State-compatible construction and separate retention carry this distinction into controlled exchanges and complete systems. Whether rechecking repairs or rolls back depends on how the reader uses source information relative to the candidate and the queried state.

AI USE STATEMENT

Generative AI assisted conceptual framing and hypothesis refinement, mathematical formulation and proof drafting/checking, feedback on experimental controls, interpretation of author-supplied results, and reference-code preparation. It also assisted literature discovery and bibliography preparation, title ideation, manuscript drafting and organization, language editing, and LaTeX, table, and figure-layout preparation. These tools generated no neural experimental measurements. The authors are responsible for the scientific conclusions, attribution, and final submission.

REPRODUCIBILITY STATEMENT

Appendices A–B provide assumptions and proofs; Appendices C–D specify implementation and evaluation. Appendix E.5 describes the fitted prediction procedure. The source package contains aggregate endpoints and mathematical and tensor-interface checks.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Jing Bi, Filippos Bellos, Junjia Guo, Yayuan Li, Chao Huang, Yunlong Tang, Luchuan Song, Susan Liang, Zhongfei Zhang, Jason J. Corso, and Chenliang Xu. When to Think and When to Look: Uncertainty-Guided Lookback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5104–5113, 2026.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. Perception Tokens Enhance Visual Reasoning in Multimodal Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3836–3845, 2025.

Mahtab Bigverdi, Linjie Li, Weikai Huang, Yiming Liu, Jaemin Cho, et al. Imaginative Perception Tokens Enhance Spatial Reasoning in Multimodal Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3D: Geometric Imagination Grounded Spatial Reasoning from Limited Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2624, 2026.

Rahul Chowdhury, Timothy A. Rupprecht, Xuan Shen, Pu Zhao, and Yanzhi Wang. ScAle: Attention Head Scaling as a Minimal Adapter for Spatial Reasoning in Vision–Language Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

Wenxiao Fan, Jingling Fu, Fang Li, Luohang Liu, Yu He, Lichen Ma, Zhiyang Yu, Weishan Bi, Junshi Huang, Yan Li, Gu Simiu, and Kan Li. Recompute or Reuse? Diagnosing and Mitigating Textual Shortcuts in VLM Self-Reflection. *arXiv preprint arXiv:2608.01930*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. doi: 10.1007/978-3-031-73337-6_9,year={2024},.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A Scalable Dataset Generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3761, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional Visual Reasoning without Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

Qingdong He, Xueqin Chen, Chaoyi Wang, Yanjie Pan, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xiangtai Li, and Jiangning Zhang. Reasoning to Edit: Hypothetical Instruction-Based Image Editing with Visual Reasoning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Jaeyun Jang, Seunghui Shin, Taeho Park, and Hyoseok Hwang. Keep it SymPL: Symbolic Projective Layout for Allocentric Spatial Reasoning in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9604–9614, 2026.

Liqiang Jing, Xiong Zhou, Siddharth Varia, Neha Anna John, Xinya Du, and Vassilis N. Ioannidis. Be Consistent! Enhancing Robust Visual Reasoning in LVLMs with Consistency Constraints. *arXiv preprint arXiv:2607.21722*, 2026.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Phillip Y. Lee, Chanho Park, Mingue Park, Seungwoo Yoo, Juil Koo, and Minhyuk Sung. Token Warping Helps MLLMs Look from Nearby Viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9570–9581, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Chaofan Ma, Zhenjie Mao, Yuhuan Yang, Fanqin Zeng, Yue Shi, Yingjie Zhou, Xiaofeng Cao, and Jiangchao Yao. Reason, Then Re-reason: Cross-view Revisiting Improves Spatial Reasoning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46534–46594, 2023.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12700–12710, 2021.

Raymond Reiter. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, 2001.

Shailaja Keyur Sampat, Akshay Kumar, Yezhou Yang, and Chitta Baral. CLEVR_HYP: A Challenge Dataset and Baselines for Visual Question Answering with Hypothetical Actions over Images. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3692–3709, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Sourabh Sharma, Sonam Gupta, and Sadbhawna. Improving Reasoning in Vision-Language Models via Perception Verified Self-Training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 17023 of *Lecture Notes in Computer Science*, pp. 334–351, 2026. doi: 10.1007/978-3-032-37592-6_19.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Shravan Venkatraman, Omkar Thawakar, Ritesh Thawkar, Abdelrahman Shaker, and Rao Muhammad Anwer. Perception Before Supervision: Self-Contained Visual Distillation from Counterfactual Blind Spots. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2026. Accepted.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Large Language Models Can Self-Correct with Key Condition Verification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12846–12867, 2024.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-Time Scaling with World Models for Spatial Reasoning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33510–33520, 2026.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 1039–1050, 2018.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking Object-Level Counterfactual Spatial Reasoning in VLMs from In-the-Wild Photos. *arXiv preprint arXiv:2607.00491*, 2026.

Shaoxiong Zhan, Yanlin Lai, Zheng Liu, Hai Lin, Shen Li, Xiaodong Cai, Zijian Lin, Wen Huang, and Hai-Tao Zheng. 3ViewSense: Spatial and Mental Perspective Reasoning from Orthographic Views in Vision-Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

APPENDIX

A Proofs of the Verification Boundaries **2**

B Response Models and State Information **4**
B.2 Unfinished edits and admissible responses 5
B.5 Query and transition sufficiency 7

C Post-State Construction and Reading **8**

D Evaluation Protocols and Task Ports **10**
D.6 Task ports and information permissions 12

E Source Verification and Joint Errors **12**
E.1 Native verification and fixed candidates 12
E.3 State-applicability controls 15
E.5 Held-out prediction and calibration 18
E.7 Dose responses and recomputation 20

F Construction and Source Retention **22**
F.1 Fixed-bank exchanges and quality–conflict effects 22
F.3 Scene banks, caches, and textual history 24

G Complete-Task Capability **29**
G.2 Larger backbones and additional families 30
G.4 Independent real-photo evaluation 33
G.5 Peer-reviewed systems and information permissions 34

H Transfer and Sequential Edits **36**
H.2 Intermediate states and recomputation 37
H.3 Source-retention policies 38

I Query and Transition Sufficiency **40**

J Computation and Cost **41**

K Limitations **43**

A Proofs of the Verification Boundaries

The main analysis asks how much correction a source can supply, when that correction reverses, and when a state change alters the answer. The proofs use the same target-relative errors throughout. Sections A.2–A.3 establish the risk and dependence boundaries; Sections A.4–A.5 then address transported evidence and query margins.

A.1 Coordinates and probability space

Propositions 1–3 concern a fixed task population or a specified conditional stratum. The candidate and evidence errors $e, z \in \mathbb{R}^p$ are random vectors on the same probability space. The evaluation matrix $W \in \mathbb{R}^{p\times p}$ is fixed, symmetric, and positive semidefinite. We assume $\mathbb{E}e^\top We < \infty$ and $\mathbb{E}z^\top Wz < \infty$. Cauchy–Schwarz then ensures that $B = \mathbb{E}e^\top Wz$ exists and $|B| \le \sqrt{AD}$. The vectors may be biased and dependent. When W is singular, equality concerns the coordinates visible through $W^{1/2}$ rather than all coordinates of the scene.

The strength g is fixed inside each expectation. A response-dependent strength requires conditional moments, as discussed in Appendix B.3. The geometric state coordinates are distinct from the native visual banks in the learned implementation. Coordinate readouts measure the specified scene variables. Held-out prediction and response-class diagnostics test whether the local account describes their changes.

A.2 Repair–rollback boundary

The risk identity exposes the correction available at small source strength and the cost of increasing that strength. Its degenerate case also specifies when the evidence cannot change any measured coordinate.

Proof of Proposition 1. The verified error is $X_g - s_1 = (1-g)e + gz$. Expanding its squared seminorm gives

$$\|X_g - s_1\|_W^2 = (1-g)^2 e^\top We + 2g(1-g)e^\top Wz + g^2 z^\top Wz. \tag{A.1}$$

Taking expectations in Eq. (A.1) yields Eq. (3). Subtracting $R(0) = A$ gives Eq. (4). The quadratic coefficient is $H = \mathbb{E}\|e-z\|_W^2 \ge 0$. For $H > 0$ and $g > 0$, $\Delta R(g) < 0$ is equivalent to $gH < 2(A-B)$, which requires $A > B$ and gives Eq. (5).

When $H = 0$, the nonnegative variable $\|e-z\|_W^2$ has zero expectation. Hence $W^{1/2}e = W^{1/2}z$ almost surely, so $A = B = D$ and every strength has risk A. □

The initial opportunity $A - B$ is not a measure of candidate quality alone. Since $z = \nu - d$ for a fixed edit displacement,

$$B = \operatorname{tr}(W \operatorname{Cov}(\nu, e)) + \mathbb{E}[e]^\top W \mathbb{E}[\nu] - \mathbb{E}[e]^\top W d. \tag{A.2}$$

Equation (A.2) separates shared perceptual fluctuations, aligned mean error, and alignment with the edit. An unbiased independent candidate and source have $B = 0$, but an under-edited candidate can have positive B even when source perception is exact. Likewise, $D = \mathbb{E}\|\nu\|_W^2 + \|d\|_W^2 - 2\mathbb{E}[\nu]^\top Wd$ for fixed d. Thus D is not pure edit magnitude, and A, B, D are related moments rather than three independent causal switches. The scalar B is an uncentered cross-moment, not a Pearson correlation coefficient.

Differentiating Eq. (4) gives the proxy-optimal strength

$$g_{\mathrm{opt}} = \operatorname{clip}_{[0,1]}\left(\frac{A-B}{H}\right), \qquad H > 0. \tag{A.3}$$

The minimizer in Eq. (A.3) and the switching point are different. Before clipping, the switch is twice the minimizer. A switch above one is outside the available mixture interval; a switch at one ties the baseline at the endpoint. Neither case supplies an observed interior crossing. This distinction motivates retaining no-crossing conditions in the prediction analysis.

A.3 SHARP DEPENDENCE REGIONS

For fixed marginal second moments, the remaining freedom is how the errors align. Bounding that alignment gives the full attainable risk interval and identifies the regions in which its sign can change.

Proof of Proposition 2. Cauchy–Schwarz gives $-\sqrt{AD} \le B \le \sqrt{AD}$. The coefficient of B in Eq. (3) is nonnegative on $[0, 1]$, so substituting its two extremes gives Eq. (6). The bounds are sharp whenever the specified moments are attainable: choose a unit vector v in the W-seminorm, a zero-mean scalar ξ with $\mathbb{E}\xi^2 = 1$, and $e = \sqrt{A}\,\xi v$, $z = \pm\sqrt{D}\,\xi v$. These constructions attain the upper and lower bounds, respectively. If $W = 0$, only $A = D = 0$ is attainable and the bounds coincide.

If $A > D$, then $(1-g)\sqrt{A} + g\sqrt{D} < \sqrt{A}$ for every $g > 0$. Its square is an upper bound on risk, proving strict improvement for every dependence. Now let $D > A > 0$ and write $a = \sqrt{A}$, $d = \sqrt{D}$. The upper bound is greater than A for every $g > 0$. The lower bound is $[a - g(a+d)]^2$, which is less than a^2 exactly when $0 < g < 2a/(a+d)$. Thus both signs are attainable throughout that interval. Above $2a/(a+d)$, even the lower bound exceeds A, so every dependence is harmful. At equality the opposed construction ties the baseline and other attainable risks are no smaller. □

The result specifies where marginal-error information is sufficient and where it is not. For $A = D > 0$, the upper bound equals A, so no dependence increases risk within the mixture interval; intermediate doses can improve risk when the errors are not perfectly aligned. At $g = 1$, only D remains. Dependence matters at intermediate strengths; complete substitution depends only on source risk.

For the analytical example in the main text, take $e \in \{-1, 1\}$ with equal probabilities and $W = 1$. Setting $z = 2e$ gives $(A, B, D) = (1, 2, 4)$ and $\Delta R = 2g + g^2 > 0$. Taking $z \in \{-2, 2\}$ independently gives $(A, B, D) = (1, 0, 4)$ and $\Delta R = -2g + 5g^2 < 0$ for $0 < g < 2/5$. The marginal distributions themselves, not just their second moments, are unchanged. A predictor restricted to those marginals cannot distinguish the two cases.

The finite-support couplings are enumerated in Table A.1. At $g = 0.2$, their risks are 1.44, 0.80, and 0.16 for shared, independent, and opposed errors. Both marginal distributions are identical. The comparison proves that a marginal-only description can fail even without estimation error.

Table A.1: **Exact couplings preserve both error distributions.** In every column $A = 1$ and $D = 4$; shared, independent, and opposed couplings have $B = 2, 0, -2$. Analytical risk and its change from the candidate are shown at each strength.

g	Shared		Independent		Opposed	
	$R(g)$	ΔR	$R(g)$	ΔR	$R(g)$	ΔR
0.000	1.000	0.000	1.000	0.000	1.000	0.000
0.100	1.210	0.210	0.850	−0.150	0.490	−0.510
0.200	1.440	0.440	0.800	−0.200	0.160	−0.840
0.400	1.960	0.960	1.000	0.000	0.040	−0.960
0.600	2.560	1.560	1.600	0.600	0.640	−0.360
0.800	3.240	2.240	2.600	1.600	1.960	0.960
1.000	4.000	3.000	4.000	3.000	4.000	3.000

Figure 2(a) displays the same finite-support couplings; panel (c) uses a separate scalar query example.

A.4 STATE TRANSPORT

Transport changes the evidence error while leaving the candidate fixed. Subtracting the two responses therefore isolates the effects of state alignment and any new construction error.

Proof of Proposition 3. Assume $\mathbb{E}\|z_T\|_W^2 < \infty$. Applying Eq. (3) to z and z_T and subtracting cancels the shared candidate term, leaving Eq. (8). For the known additive edit $T(s) = s + d$, the source error relative to s_1 is $z = \nu - d$ and the transported error is $z_T = \nu$. With fixed d and

zero-mean ν, $D - D_T = \|d\|_W^2$. Independent, zero-mean e, ν give $B = B_T = 0$, completing the stated special case. □

For a known affine transformation $T(s) = Js + b$, source noise becomes $J\nu$. Exact transport aligns the state semantics but can amplify uncertain coordinates. A learned transformation adds its own construction error. Consequently, the general moment difference in Eq. (8), rather than the additive special case, determines whether transformed evidence helps. Noninvertible operations such as removal can still be applied forward; they do not require an inverse source likelihood.

A.5 Answer-margin bridge

An answer changes when a query margin crosses zero, which need not coincide with a return to the initial coordinate risk. The affine case makes this distinction explicit before extending it to multiple choices.

Lemma A.1 (Query crossing)**.** *For a binary threshold at b, let $x > b$, $o < b$, $c = x - b$, and $a = b - o$. Along $x_g = (1-g)x + go$, the signed margin is Eq. (7). It is positive below $c/(a+c)$ and negative above it; the equality case follows the query's tie rule.*

Proof. Subtract b to obtain $(1-g)c - ga$. Since $a + c > 0$, this is strictly decreasing and has its unique zero at $c/(a+c)$. □

Lemma A.1 extends pairwise to affine multiclass scores. The first crossed margin between a_1 and a competitor determines loss of correctness. Selecting a_0 additionally requires its score to exceed every other competitor. A state-risk increase therefore need not produce an answer change, and an answer change need not be rollback. Table 2 predicts answer correctness, while the native transition census in Table E.1 separately identifies old-answer return.

Figure 2(c) illustrates the distinction: for scalar errors $e = 1, z = -2$, the risk minimum is $1/3$, an equal-margin query crosses at $1/2$, and risk returns to its initial value at $2/3$.

Table A.2 evaluates the position, risk, and margin along this analytical path.

Table A.2: **Analytical separation of the risk optimum, answer crossing, and return to candidate risk.** The underlying path is $x_g = 1 - 3g$, target zero, and query boundary $-1/2$; decimals display exact rational constructions.

g	**Position** x_g	**Risk** $R(g)$	**Margin** $m(g)$
0.000	1.000	1.000	1.500
0.200	0.400	0.160	0.900
0.333	0.000	0.000	0.500
0.500	−0.500	0.250	0.000
0.667	−1.000	1.000	−0.500
1.000	−2.000	4.000	−1.500

B Response Models and State Information

A local mixture becomes an explanation only within a specified response class. Sections B.1–B.4 examine estimated moments, unfinished edits, adaptive strengths, and empirical departures from that class. Sections B.5–B.6 ask a different question: what information and reader regularity are needed to turn a state into a reliable answer?

B.1 Moment and response errors

Write $\Delta = R(g) - R(0)$ for the coordinate-risk change. A boundary is most sensitive where the predicted effect is close to zero. Lemma B.1 compares moment-estimation error with distance to the switching boundary; Lemma B.2 then bounds departures from the assumed response.

Lemma B.1 (Moment perturbation). *Fix $g \in [0,1]$. If $|A - \widehat{A}| \le \epsilon_A$, $|B - \widehat{B}| \le \epsilon_B$, and $|D - \widehat{D}| \le \epsilon_D$, then*

$$|\Delta - \widehat{\Delta}| \le (2g - g^2)\epsilon_A + 2g(1-g)\epsilon_B + g^2\epsilon_D. \tag{B.1}$$

The predicted sign is preserved whenever its absolute value exceeds this bound.

Proof. Write $\Delta = (g^2 - 2g)A + 2g(1-g)B + g^2 D$, subtract the estimated expression, and apply the triangle inequality. On $[0,1]$, the absolute coefficient of A is $2g - g^2$ and the other coefficients are nonnegative. A perturbation smaller than the distance to zero cannot reverse a sign. □

Equation (B.1) distinguishes uncertainty in the inputs from a small predicted effect. The bound uses separate bounds on moment-estimation errors and gives a deterministic sign certificate.

Lemma B.2 (Residual response). *Suppose the actual state readout is $Y_g = (1-g)X + gO + r_g$, the baseline is $Y_0 = X$, and $\mathbb{E}\|r_g\|_W^2 \le \varepsilon_g$. Let $R(g)$ denote the mixture risk. Then*

$$\left|\mathbb{E}\|Y_g - s_1\|_W^2 - R(g)\right| \le 2\sqrt{R(g)\varepsilon_g} + \varepsilon_g. \tag{B.2}$$

If $|R(g) - A|$ exceeds this bound, the mixture and actual risk changes have the same sign.

Proof. Set $v_g = (1-g)e + gz$. The difference is $2\mathbb{E}[v_g^\top W r_g] + \mathbb{E}\|r_g\|_W^2$. Cauchy–Schwarz bounds the cross-term by $2\sqrt{R(g)\varepsilon_g}$, proving Eq. (B.2). Subtracting the shared baseline A gives the sign claim. □

Lemma B.2 gives the out-of-plane check a quantitative interpretation. It separates two explanations for changed transfer behavior: a different effective strength and a response outside the mixture class. Applying this certificate requires a bound on the residual moment in addition to the aggregate prediction scores.

B.2 Collinear unfinished edits

An exact partial edit leaves the candidate between the source and the target. Averaging it with the source moves in the wrong direction, even when source perception is perfect.

Lemma B.3 (Collinear under-edit exclusion). *Let $s_1 = s_0 + d$, $O = s_0$, and $X = s_1 - \alpha d$ with $0 < \alpha < 1$ and $\|d\|_W^2 > 0$. Every $g \in (0,1]$ increases state error under raw-state mixing. Replacing the source by the correctly transformed $O_T = s_1$ strictly decreases it.*

Proof. Raw mixing gives $X_g - s_1 = -[\alpha + g(1-\alpha)]d$, whose squared norm exceeds $\alpha^2\|d\|_W^2$. With $O_T = s_1$, the error is $-(1-g)\alpha d$ and is strictly smaller for every positive $g \le 1$. □

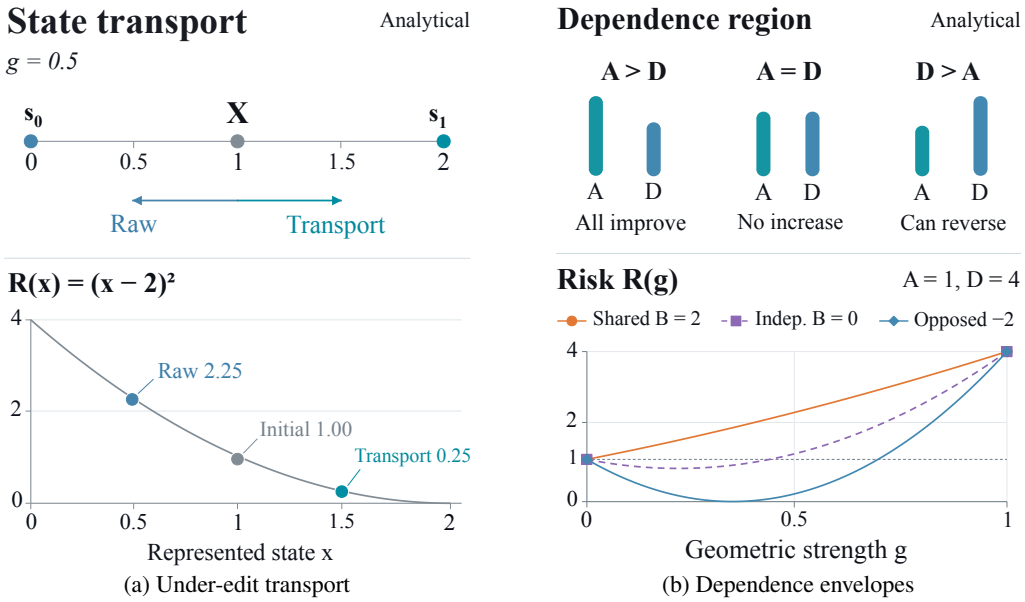
Lemma B.3 excludes raw averaging for the exact collinear condition it defines. There, mixing with the original state cannot supply the missing part of the edit; correction requires a transformed source or another response class. Natural candidates can additionally contain perceptual and noncollinear errors, so the exclusion does not classify every real-photo repair. Figure B.1 depicts this geometric restriction. The result identifies the missing transformation that a successful correction must supply.

B.3 Adaptive evidence strengths

A geometric evidence-use policy can select its strength from the current example. The fixed-strength identity then has to be conditioned on the variables used for selection before it is averaged over examples.

For a random strength G correlated with state errors, risk is

$$\mathbb{E}\left[(1-G)^2\|e\|_W^2 + 2G(1-G)e^\top W z + G^2\|z\|_W^2\right]. \tag{B.3}$$



(a) Under-edit transport (b) Dependence envelopes

Figure B.1: **Raw-state mixing has geometric and statistical limits.** (a) Transport moves an unfinished edit toward its target; the raw source moves it away. (b) Sharp envelopes separate uniform improvement from dependence-sensitive responses. The analytical curves hold $A = 1$ and $D = 4$ fixed while changing B.

Equation (B.3) generally differs from $R(\mathbb{E}G)$. Even independence does not remove the second moment of G. If G is independent of (e, z), then $\mathbb{E}R(G) = R(\mathbb{E}G) + H\,\mathrm{Var}(G)$. This identity follows by taking expectations of the quadratic in Eq. (4). For adaptive policies, conditioning on the variables that select G restores the fixed-strength analysis within each stratum; averaging the resulting conditional risks is valid, whereas substituting global moments need not be. The reported development calibration is therefore evaluated on held-out conditions rather than justified by a universal average-strength assumption.

B.4 Zero dose and response residuals

The theoretical reference is candidate X, whereas the empirical zero-dose reference is $Y_{\lambda=0}$. Masking direct source keys does not remove historical text or secondary injections, and Eq. (E.1) need not give $g_\theta(0, h) = 0$. For fixed strengths $g, g_b \in [0, 1]$, subtraction of the quadratic gives

$$R(g) - R(g_b) = (g - g_b)\{H(g + g_b) - 2(A - B)\}. \tag{B.4}$$

By Eq. (B.4), for $H > 0$ and $g > g_b$, improvement relative to that baseline requires $g + g_b < 2(A - B)/H$. The nontrivial return occurs at $g = g_{\text{switch}} - g_b$ when it lies in the tested interval. This identity follows by factoring $g^2 - g_b^2$ and does not identify g_b from a plotted dose axis.

Let $v = O - X$ and $\|v\|_W^2 > 0$. The unrestricted line projection and constrained segment projection are

$$\gamma_\lambda = \frac{\langle Y_\lambda - X, v\rangle_W}{\|v\|_W^2}, \qquad \widetilde{g}_\lambda = \mathrm{clip}_{[0,1]}(\gamma_\lambda). \tag{B.5}$$

Using the projections in Eq. (B.5), define $r_\perp = Y_\lambda - X - \gamma_\lambda v$ and $r_{\text{seg}} = Y_\lambda - X - \widetilde{g}_\lambda v$. Orthogonality gives

$$\|r_{\text{seg}}\|_W^2 = \|r_\perp\|_W^2 + (\gamma_\lambda - \widetilde{g}_\lambda)^2 \|v\|_W^2. \tag{B.6}$$

Equation (B.6) shows that a clipped residual is not necessarily perpendicular. Figure E.3 illustrates a response beyond the source endpoint and off the line. Zero separation leaves the coefficient unidentified in the measured coordinates.

For an actual response $Y_\lambda = X_{g_\lambda} + r_\lambda$, Lemma B.2 applies to both the intervention and the baseline. If $C_\lambda = 2\sqrt{R(g_\lambda)\varepsilon_\lambda} + \varepsilon_\lambda$, then

$$|(R_\lambda - R_0) - \{R(g_\lambda) - R(g_b)\}| \le C_\lambda + C_0. \tag{B.7}$$

The triangle inequality proves Eq. (B.7). A small intervention residual does not compensate for a large baseline residual. Adaptive strengths require conditional moments before averaging, as in Appendix B.3.

B.5 State semantics and sufficiency

State compatibility is a statement about the world to which the evidence applies. Expressing the target as a transformation of source uncertainty clarifies both historical correspondence and information that the inputs cannot reveal.

Let $\Pi(ds \mid I_0)$ be uncertainty about the original scene and $f(a \mid s, q)$ an answer kernel. The counterfactual target is

$$P(a \mid I_0, u, q) = \int f(a \mid T_u(s), q)\, \Pi(ds \mid I_0). \tag{B.8}$$

Equation (B.8) expresses the edited-state distribution as the pushforward of the original uncertainty. For a deterministic query it is equivalently the composed query $q \circ T_u$ on the source. For invertible edits, a source likelihood in post-state coordinates uses T_u^{-1}; for removal, the forward integral remains valid without an inverse. Reusing the same photograph does not create a second independent observation of an edited world.

Geometric correctness does not guarantee query sufficiency. A question about the object that was originally leftmost requires a historical referent. If two visually indistinguishable objects change places, a post-state that omits identity correspondence can be accurate yet insufficient. Retaining a state-indexed identity mapping or evaluating the composed query on the original scene preserves the missing relation. Historical correspondence is illustrated and tested behaviorally in Appendix I.

Lemma B.4 (Indistinguishable public inputs)**.** *If two equally probable scenes have identical allowed inputs but different correct counterfactual answers, every predictor of those inputs has average error at least $1/2$ on that two-scene population.*

Proof. The output distribution is the same on both scenes. For their distinct correct answers a, b, average success is $[P(a) + P(b)]/2 \le 1/2$. □

Lemma B.4 concerns an explicit information ambiguity, not the measured prevalence of such scenes in either benchmark. A deterministic constructor cannot resolve the missing information. Even minimizing representation MSE may average incompatible possibilities: with target Y and non-negative weight w, the conditional minimizer is $\mathbb{E}[wY \mid x]/\mathbb{E}[w \mid x]$ when the denominator is positive, obtained by differentiating the conditional quadratic. It is a prediction objective rather than a guarantee of a physically realizable scene.

B.6 Native-state answer stability

Native visual-bank accuracy and answer accuracy are connected through the reader. A local regularity condition gives a sufficient representation-error bound for preserving an answer, while keeping the reader and its non-scene inputs fixed.

Lemma B.5 (Local logit stability)**.** *Fix the reader and all non-scene inputs. Suppose the correct logit at Z_1 exceeds each competitor by at least $\gamma > 0$, and every logit is L-Lipschitz on the neighborhood containing Z_1 and $\widehat{Z}$, with $L > 0$. If $\|\widehat{Z} - Z_1\| < \gamma/(2L)$, the answer is preserved.*

Proof. The correct logit decreases by at most $L\|\widehat{Z} - Z_1\|$ and any competitor increases by at most the same amount. Every margin remains at least $\gamma - 2L\|\widehat{Z} - Z_1\| > 0$. □

Lemma B.5 requires a local constant, a positive reference margin, and a fixed reader. For the evaluated readers, representation error measures construction quality and answer accuracy tests actual use; a general link between them additionally needs the stated local regularity. Exact replay from identical complete banks, positions, masks, public option banks, and empty caches is a separate deterministic interface identity. Keeping an old secondary injection breaks that identity even when an initial bank has been replaced.

An object's pivot orbit obeys $p'_A = p_B + R_z(\theta)(p_A - p_B)$, whereas an in-place spin changes orientation while preserving the center. Independent absolute translations commute; relative translations that read a moving reference object may not. Removal has no universal inverse. Figure H.1 illustrates why the implementation preserves these operation distinctions rather than enforcing an unrestricted cycle-consistency rule.

C Post-State Construction and Reading

A request must preserve its edit semantics before any representation is constructed. The implementation therefore separates parsing, scene-bank construction, and reading. The specifications below follow that path, then give training settings and the checks needed for a complete exchange of scene information.

C.1 Separating edits from terminal queries

The public interface supplies the original images, unchanged question, and original answer options. Image roles distinguish editable scene images from answer candidates and analogy context. The constructor receives only editable scene banks and a self-contained edit instruction. Terminal queries, answer choices, correct-option identities, test-time post-edit images, and private geometry are excluded from construction.

A frozen original language-model checkpoint performs text-only parsing with greedy decoding and a 384-token output limit. Its output contains a task type, an ordered list of at most two edits, and the terminal question with the edit prefix removed. Object descriptions, pivots, angles, units, frames, and orientation conventions remain literal. Answer options are preserved externally. The parser instruction is:

> Separate a hypothetical geometric edit from the question about its result. Do not answer or infer geometry. Preserve exact object descriptions, pivots, units, reference frames, image roles and answer options. Distinguish rotating an object around another object from spinning it in place. Copy each edit as self-contained text. Do not transform answer-option or analogy images. Return only the specified JSON. Use unsupported when safe separation is not possible, when historical identity would be lost, or when the request needs physical dynamics or more than two edits. Never invent missing information.

Validation checks task type, a unique separation, preservation of parameters and referents, supported image roles, and unchanged option mapping. An invalid parse, unsupported montage, unsafe historical reference, dynamics request, or excessive edit length invokes native fallback. Fallback uses the same trained reader with the original images and original question; it does not switch to a different base model. Parsing work and fallback outcomes remain part of the declared system. Manually separated questions are diagnostic comparisons, not the automatic endpoint.

C.2 Native banks and the constructor

A visual bank consists of a tensor $Z^k \in \mathbb{R}^{n_s \times N \times D_k}$ for n_s sources, together with valid-token masks, scene masks, image roles, native positions, and injection-point information. The constructor assumes corresponding spatial positions across its input banks. An incompatible layout follows native fallback. Qwen3-VL includes initial image embeddings and secondary DeepStack entries; Gemma uses its native projector interface (Bai et al., 2025; Gemma Team, 2025). Complete replacement applies at every scene-related entry actually present.

Each bank is projected to width 512. The projected banks are summed and divided by $\sqrt{K}$, where K is the number of banks. Token centers use two-dimensional coordinates in $[-1, 1]$. Sine and cosine features at frequencies $\pi 2^j$, $j = 0, \ldots, 7$, give a 32-dimensional position code, followed by a learned projection. One of three scene-view embeddings is added. Public answer-candidate tokens are excluded from constructor self-attention, and padded positions are zeroed after each block.

The four pre-layer-normalized blocks each contain eight-head visual self-attention, cross-attention to the edit instruction, and a width-2048 GELU feed-forward layer with residual connections. Dropout is zero. A final normalization precedes bank-specific output projections. Their zero-initialized weights

and biases make Eq. (9) an identity at initialization. An explicit no-op branch returns its input exactly after training as well.

Instruction features come from the frozen original language model's final text hidden states, with reader adapters disabled. The edit limit is 256 tokens; an overlong edit falls back instead of truncating a pivot or unit. Sequential edits reuse the constructor on its predicted intermediate banks without receiving a true intermediate image. Original banks remain immutable. Query reuse holds the extracted edit fixed, since the parser reads the original question.

C.3 State targets and answer objectives

The frozen visual encoder supplies before, intermediate, and final target banks. For bank k, a training-set RMS scale s_k, lower-bounded by 10^{-6}, is held fixed. Token weights balance whole-scene fidelity with edit-related change,

$$d_i^k = \frac{\|Z_{1i}^k - Z_{0i}^k\|}{\sqrt{D_k}}, \qquad w_i^k = \frac{1}{2N} + \frac{d_i^k}{2\sum_j d_j^k}. \tag{C.1}$$

For zero total change, the weight is $1/N$. Targets, weights in Eq. (C.1), and scales are detached. Averaging over sources and banks gives

$$\mathcal{L}_{\text{state}} = \frac{1}{n_{\text{s}}} \sum_x \frac{1}{K} \sum_k \sum_i w_{xi}^k \frac{\|\widehat{Z}_{xi}^k - \text{sg}\, Z_{1,xi}^k\|^2}{D_k s_k^2}. \tag{C.2}$$

Uniform weighting replaces w_i^k with $1/N$ without changing Eq. (C.2) otherwise. With two edits, intermediate and final state losses receive equal weight and the final gradient passes through the first predicted edit. State loss is computed once per source, not once per terminal query.

The three answer terms in Eq. (10) use predicted post-state banks with a counterfactual query, true training post-state banks with that query, and original banks with an original-scene factual query. Answer-token loss is averaged within query, across the available queries of a source, and then across sources. The answer target contains the answer inside `<final>` delimiters and a termination token, without a ground-truth reasoning trace. The query schedule selects a changed and an invariant question when both exist, otherwise the next available valid question. Losses are averaged over the questions actually used. Freezing the language head still permits answer gradients to reach the constructor and reader adapters.

C.4 Optimization and inference

The optimization recipe uses a logical batch of four sources, two available queries per source, and single-source/query microbatches. The constructor learning rate is 10^{-4}. Reader LoRA uses learning rate 10^{-5}, rank 16, scale 32, and zero dropout (Hu et al., 2022). Adapters target the language attention and feed-forward linear projections; vision, projector, embeddings, and language-head parameters remain frozen. AdamW (Loshchilov & Hutter, 2019) uses momentum coefficients $(0.9, 0.95)$, numerical epsilon 10^{-8}, weight decay 0.01, and global gradient clipping at one. Backbone operations use BF16; losses use FP32.

The prescribed curriculum has 128 constructor-only state-loss updates followed by joint training, with separate warmup controls described in Appendix F.4. Sixteen warmup updates precede a constant learning rate. The default state coefficient is $\beta = 1$. The development grid for append and replacement combines reader rates $\{3\times10^{-6}, 10^{-5}\}$ with $\beta \in \{0.3, 1, 3\}$; their shared configuration maximizes the mean development score of the two arms, breaking ties toward smaller β and then smaller rate. Paired SFT varies reader rate over $\{3\times10^{-6}, 10^{-5}, 3\times10^{-5}\}$ and weight decay over $\{0, 0.01\}$.

Training images use a 448-pixel training resolution with native family processing. The Qwen scene-pixel budget is 1280×32^2; optional Gemma pan-and-scan is disabled. The context limit is 8192 tokens. Terminal reading uses greedy decoding and at most 512 new tokens except in explicit token-cap comparisons. An over-budget request follows fallback.

C.5 Complete scene-bank exchange

Append supplies labeled original and predicted scene banks but does not duplicate option images. Replacement supplies the predicted banks at every scene-related injection point. Native positions, masks, special tokens, and rotary-position metadata are reconstructed for the actual reader input. Both modes clear attention caches and visual-injection buffers before terminal reading. Retaining an old secondary bank is the intentional fault condition in Table F.4.

D Evaluation Protocols and Task Ports

The same accuracy difference can answer different questions depending on what is fixed. Native transitions measure verification behavior; controlled descriptions and coordinates identify conditional responses; fixed-bank exchanges isolate source retention; separately trained systems measure capability. This section defines the populations and measurements before specifying what each adapted comparator receives.

D.1 Tasks and data separation

MindEdit-L4 supplies real-photo object-level edits with the original questions, multiview images, answer options, and native scoring (Yu et al., 2026). STARE-rigid-text selects the public 3D text-instructed rotation and translation tasks by task taxonomy (Li et al., 2026). Public answer images remain available to every method. An extra post-edit or intermediate image is an oracle unless the public task itself supplies it. The independent real-photo cohort is evaluated separately from MindEdit, not added to its test set. Rigid-edit accuracy and full-benchmark results have separate populations and retain their native metrics.

Controlled before/after rendering uses analytic objects and scanned household assets (Downs et al., 2022; Greff et al., 2022). This follows the use of explicit scene structure for compositional visual reasoning (Johnson et al., 2017; Yi et al., 2018), but separates rigid state changes from the collision dynamics studied in CLEVRER (Yi et al., 2020). A source scene is the split unit, including all its views, edits, questions, paraphrases, and intermediate renders. Method selection and predictor calibration use separate development roles. Asset, camera, background, composition, and operation shifts are evaluated independently. These rendered controls are distinguished from the natural-photo evaluations.

The geometric frame is right-handed with vertical z. Translation changes an object's center. A pivot orbit uses the reference object's current center; in-place spin preserves the center. Observer changes move the queried frame rather than the scene objects. Removal is not assigned a universal inverse. Relative translations that depend on moving objects need not commute. Cases with unidentified metric scale, an undefined intrinsic front, an unsafe historical referent, or physical dynamics do not acquire missing information from a parser.

D.2 Evidence and candidate construction

Ordinary verification generates an answer and then performs one source-based check, without a supplied correct description. Replay extracts an observable candidate and changes the verification context. Correct, under-edited, wrong-pivot, and unobserved or unjudgeable descriptions remain separate. The replay average is therefore not a weighted reconstruction of the ordinary-verification transition table.

The supplied-card experiment holds the source, edit, candidate description, and terminal query fixed within an evidence contrast. Candidate descriptions are constructed without the terminal query or options and are checked separately for geometry and query sufficiency. Each is described as fallible without revealing its correctness. Old-image, old-text, neutral, no-extra-image, unrelated-text, and independently supplied post-state-text conditions change the evidence, not the candidate. Independent post-state text tests the relevance of the target state; it is not the output of the deployed constructor.

The shared state-explicit instruction is:

> The images labeled ORIGINAL show the scene before the hypothetical edit. The candidate description refers to the scene after the edit and may contain an error. Check it using the available evidence and the stated edit. An original image is not a new observation of the edited scene. Answer the question about the edited scene, preserving facts that the edit does not change.

The before-state control changes the requested target consistently while retaining candidate labels relative to the post-state. Evidence order, message role, padding, and weaker rechecking language are separate interventions. Diagram and coordinate-table controls match informational content more closely than a natural photograph and a short text description can.

D.3 Source-reading dose

The intervention adds $\log \lambda$ to attention scores from newly generated response positions to originally visible scene-image keys. At $\lambda = 0$, those keys are masked; $\lambda = 1$ is the identity intervention. Public option keys, originally invisible positions, and prior key/value tensors are unchanged. Historical text and secondary injections can retain source information. Thus zero dose removes one direct reading route rather than all knowledge of the source.

The recorded dose profiles use 0, 0.25, 0.5, 1, 2, and 4. Intermediate markers at 0.125, 0.375, 0.75, 1.5, and 3 are linear interpolants for visualization. The response coefficient g is calibrated from development readouts, not set to λ. Blur preserves image dimensions; changing resolution also changes visual-token counts, so they are not pooled as the same intervention. Evidence before construction and evidence after a candidate test different opportunities for using the source.

D.4 Outcomes and statistical units

For correctness indicators $Y^{\mathrm{pre}}, Y^{\mathrm{post}}$, repair is $P(Y^{\mathrm{pre}} = 0, Y^{\mathrm{post}} = 1)$ and damage is $P(Y^{\mathrm{pre}} = 1, Y^{\mathrm{post}} = 0)$ over all requests. Their difference equals the accuracy change. Damage is partitioned into confirmed return to a_0, another identified error, and an error for which a_0 is unidentified. An omitted old option or a geometrically similar wrong option is not sufficient to label rollback.

Card outcomes are new-answer marginal distributions. A rise in old-answer share is not a reconstructed count of correct-to-old transitions. Fixed-bank comparisons change the input-use rule with the constructor and reader fixed. Trained-system comparisons also include learning and cannot assign the total gain to prevented rollback. Native capability retains unsupported requests through fallback to the same trained reader.

A source scene clusters its edits, views, paraphrases, and repeated decodes. A training run, not an output token or query, is the replication unit for learned-system variability. The reported accuracy values are point estimates. Eligibility counts in the prediction experiments are actual scoring denominators; repeated readouts are not additional training seeds. Individually rounded components can differ from displayed totals. Percentage-point differences are formed before display rounding.

D.5 Shared-training comparisons

Paired SFT uses original-image counterfactual questions, training post-state images with terminal questions, and original-scene factual questions, with the three answer weights in Eq. (10). It has no constructor. Labeled append and replacement use the same constructor family, state objective, parser, and answer objectives, but train their own weights. Append supplies labeled original and predicted scene banks and includes public answer images only once. The fixed-bank comparisons are needed precisely because an end-to-end comparison does not hold those learned weights fixed.

The answer-only capacity control keeps the constructor and replacement interface while removing state supervision. Table F.6 additionally separates that objective from warmup. Outcome GRPO (Shao et al., 2024) uses correctness rewards with four sampled responses per source, temperature one, a 512-token cap, ratio clipping to $[0.8, 1.2]$, and a reference-policy KL coefficient of 0.01. Extra policy samples are computation, not additional training examples with independent scene content.

The paired-data text arm predicts a post-state description without extra geometric labels. Geometry-supervised text additionally receives geometric-card token supervision, followed by an empty-cache forward pass without source scene images. Its recipe gives equal weight to card and card-conditioned answer losses, uses a 384-token card cap, and mixes 80% true training cards with 20% cards from the initial model. All test cards are predicted. The comparison marks its additional geometric labels as P+G.

D.6 Task-port implementation and information

Section 2.1 positions the evidence-use question relative to prior methods. Here we specify their task adaptations. Mirage interleaves latent visual tokens with text and learns visual targets before task supervision (Yang et al., 2026). Its same-data port retains four latent positions and two supervised phases, with a cosine visual-target term weighted by 0.1 in the first phase. Fixed-bank exchanges separately hold construction constant to identify retention.

PVST verifies perception and regenerates reasoning conditioned on accepted captions (Sharma et al., 2026). Its port uses original-scene captions and applies the hypothetical edit in reasoning. PaddleOCR, all-MiniLM-L6-v2, and FG-CLIP-large filter training examples, defining the additional-information category P+F. CVPD curates perception-oriented counterfactual blind spots for distillation (Venkatraman et al., 2026); its port uses the shared paired data. Table 3 gives the originating venues and reports evaluations under the present task protocol.

Hypothetical scene edits differ from the counterfactual interventions used to remove language bias in VQA (Niu et al., 2021). Public questions and options remain unchanged across ports. Text, geometry, generated views, and latent banks can each carry a transformed state; their additional labels, expert calls, and test-time resources are recorded in Appendix G.5. These permissions distinguish a representation comparison from a comparison that supplies more spatial information.

E Source Verification and Joint Errors

A positive average verification effect can conceal obsolete-state returns. The outcome analyses in Sections E.1–E.2 identify those returns and their candidate dependence. Controls then test state applicability, while the coordinate and prediction studies ask whether joint errors explain responses beyond the conditions used to fit them.

E.1 Native verification and candidate replay

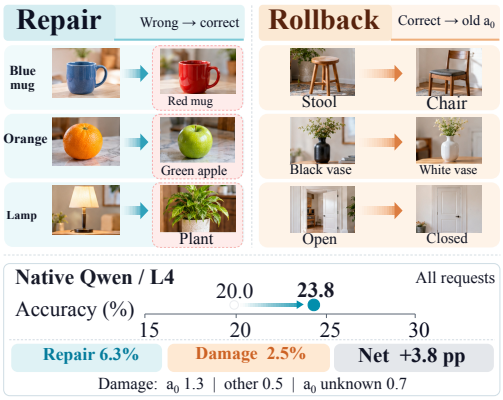
Native verification can improve its average score while undoing correct edits. All four systems in Table E.1 exhibit this combination. For Qwen3-VL-4B, 6.3% repair and 2.5% damage yield a 3.8-point gain, yet 1.3% of all requests return to an identified pre-edit answer.

Table E.1: **Native verification on MindEdit-L4.** Accuracy and transition masses use all requests (%); net change is in pp. The final three columns partition damage. Spatial-SSRL uses its 4B reference configuration.

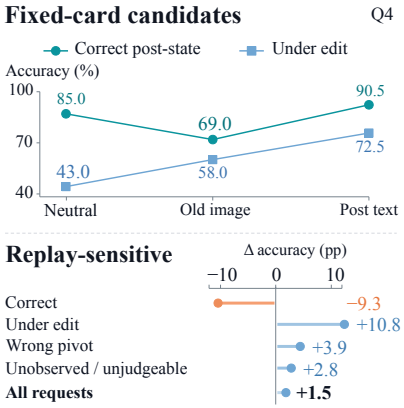
Model	Accuracy (%) ↑		Transitions (%)		Damage partition (%)			Net
	Before	After	Repair	Damage	Old a_0	Other	Unk.	
Qwen3-VL-4B	20.0	23.8	6.3	2.5	1.3	0.5	0.7	3.8
Gemma-3-4B	16.0	19.4	5.1	1.7	0.7	0.5	0.4	3.4
Qwen3-VL-8B	24.7	28.8	6.2	2.2	1.1	0.6	0.6	4.1
Spatial-SSRL (4B reference)	29.0	32.4	5.3	1.9	0.7	0.7	0.5	3.4

Table E.2 instead fixes an observable endogenous description for replay. Correct descriptions have negative source effects, whereas under-edited and wrong-pivot descriptions have positive effects. The smaller all-request effect includes unobserved or unjudgeable states rather than treating them as incorrect. The effect reverses with the quality of the observable description. Figure E.1 places the native census beside candidate replay. Panel (a) measures paired transitions from an existing answer;

panel (b) changes the evidence supplied to a fixed description. The two comparisons link a native failure to a controlled candidate-quality interaction.



(a) Native verification



(b) Fixed cards and replay

Figure E.1: **Native transitions and fixed candidates reveal complementary source effects.** (a) The all-request census separates repair from damage and confirmed old-answer return. (b) Fixed-card accuracies and endogenous replay show candidate dependence on their respective populations. Photographs illustrate the transitions rather than report individual model outputs.

Table E.2: **Endogenous-candidate replay.** Accuracy changes (pp) under source evidence, grouped by observable candidate state. Replay and fixed cards use different populations.

Candidate stratum	Accuracy change (pp)	
	Qwen3-VL-4B	**Gemma-3-4B**
Correct	−9.3	−5.5
Under-edited	10.8	6.1
Wrong pivot	3.9	2.3
Unobserved or unjudgeable	2.8	1.6
All requests	1.5	1.1

E.2 Fixed-description outcomes and interactions

Accuracy and old-answer attraction describe different parts of the same answer distribution. For unfinished-edit Qwen descriptions, source images raise correct-answer accuracy from 43.0% to 58.0% and old-answer selection from 21.9% to 28.0%. Other errors fall enough for both changes to occur. Improved accuracy can therefore coexist with stronger attraction to obsolete answers.

Tables E.3–E.5 place absolute outcomes beside their matched-neutral effects for each checkpoint. These columns are complementary summaries of one fixed-description experiment. Old images harm correct descriptions and help unfinished ones; independently supplied post-state text improves accuracy and reduces old-answer selection across candidate classes. No-extra-image and unrelated-text conditions stay near neutral. The relevant contrast is state-bearing content, rather than additional context alone.

Gemma retains the sign interaction at different magnitudes. Its wrong-pivot descriptions gain 4.0 accuracy points from old images while old-answer selection rises by 20.2 points. A source can correct one kind of error while concentrating the remaining errors on an obsolete relation. The larger Qwen checkpoint improves several absolute endpoints without eliminating this interaction.

Table E.6 subtracts the correct-description effect from the unfinished-edit effect within that same experiment. Old-image interactions are 31.0 points for Qwen-4B, 19.2 for Gemma, and 22.9 for Qwen-8B. Candidate state changes source value across all three checkpoints.

Table E.3: **Qwen3-VL-4B: fixed-description outcomes and evidence effects.** Answers partition into correct, old, other, and invalid outcomes (%). The last columns give accuracy and old-answer changes from neutral (pp), computed before display rounding. Post-state text is supplied independently.

Description	Evidence	Answer distribution (%)				Change (pp)	
		Correct	**Old a_0**	**Other**	**Invalid**	**ΔAcc.**	**ΔOld**
Correct	Neutral	85.0	5.5	8.7	0.8	–	–
	Old image	69.0	20.5	9.7	0.8	−16.0	15.0
	Old text	74.5	15.1	9.6	0.8	−10.5	9.5
	Post-state text	90.5	1.4	7.3	0.8	5.5	−4.1
	No extra image	85.8	5.1	8.3	0.8	0.8	−0.4
	Unrelated text	85.2	5.2	8.8	0.8	0.2	−0.4
Under-edited	Neutral	43.0	21.9	34.3	0.8	–	–
	Old image	58.0	28.0	13.2	0.8	15.0	6.1
	Old text	55.2	26.8	17.2	0.8	12.2	4.9
	Post-state text	72.5	4.3	22.4	0.8	29.5	−17.6
	No extra image	42.5	21.5	35.2	0.8	−0.5	−0.4
	Unrelated text	42.4	21.0	35.8	0.8	−0.6	−0.9
Wrong pivot	Neutral	25.5	28.7	45.0	0.8	–	–
	Old image	32.0	45.7	21.5	0.8	6.5	17.0
	Old text	30.6	41.8	26.8	0.8	5.1	13.1
	Post-state text	55.5	7.0	36.7	0.8	30.0	−21.8
	No extra image	25.2	28.1	45.9	0.8	−0.3	−0.6
	Unrelated text	25.8	27.2	46.2	0.8	0.3	−1.6

Table E.4: **Gemma-3-4B: fixed-description outcomes and evidence effects.** Answers partition into correct, old, other, and invalid outcomes (%). The last columns give accuracy and old-answer changes from neutral (pp), computed before display rounding. Post-state text is supplied independently.

Description	Evidence	Answer distribution (%)				Change (pp)	
		Correct	**Old a_0**	**Other**	**Invalid**	**ΔAcc.**	**ΔOld**
Correct	Neutral	79.2	7.6	12.0	1.2	–	–
	Old image	69.3	20.1	9.4	1.2	−9.9	12.4
	Old text	72.7	15.9	10.2	1.2	−6.5	8.3
	Post-state text	82.6	2.6	13.6	1.2	3.4	−5.1
	No extra image	79.7	7.3	11.8	1.2	0.5	−0.4
	Unrelated text	79.3	7.2	12.3	1.2	0.1	−0.4
Under-edited	Neutral	37.2	24.0	37.6	1.2	–	–
	Old image	46.5	35.6	16.7	1.2	9.3	11.5
	Old text	44.8	33.0	21.1	1.2	7.6	8.9
	Post-state text	55.5	6.9	36.4	1.2	18.3	−17.1
	No extra image	36.9	23.5	38.4	1.2	−0.3	−0.5
	Unrelated text	36.8	22.9	39.0	1.2	−0.4	−1.1
Wrong pivot	Neutral	19.7	30.8	48.3	1.2	–	–
	Old image	23.7	51.0	24.0	1.2	4.0	20.2
	Old text	22.9	46.3	29.6	1.2	3.2	15.5
	Post-state text	38.3	9.7	50.8	1.2	18.6	−21.2
	No extra image	19.5	30.1	49.2	1.2	−0.2	−0.7
	Unrelated text	19.9	29.2	49.7	1.2	0.2	−1.7

Table E.5: **Qwen3-VL-8B: fixed-description outcomes and evidence effects.** Answers partition into correct, old, other, and invalid outcomes (%). The last columns give accuracy and old-answer changes from neutral (pp), computed before display rounding. Post-state text is supplied independently.

Description	Evidence	Answer distribution (%)				Change (pp)	
		Correct	Old a_0	Other	Invalid	ΔAcc.	ΔOld
Correct	Neutral	88.5	4.2	6.6	0.7	–	–
	Old image	76.7	15.4	7.2	0.7	−11.8	11.2
	Old text	80.7	11.3	7.2	0.7	−7.8	7.1
	Post-state text	92.6	1.1	5.7	0.7	4.1	−3.1
	No extra image	89.1	3.9	6.3	0.7	0.6	−0.3
	Unrelated text	88.6	3.9	6.7	0.7	0.1	−0.3
Under-edited	Neutral	46.5	20.6	32.2	0.7	–	–
	Old image	57.6	28.4	13.3	0.7	11.1	7.8
	Old text	55.5	26.7	17.1	0.7	9.0	6.1
	Post-state text	68.3	5.0	26.0	0.7	21.8	−15.6
	No extra image	46.1	20.2	33.0	0.7	−0.4	−0.4
	Unrelated text	46.1	19.7	33.5	0.7	−0.4	−0.9
Wrong pivot	Neutral	29.0	27.4	42.9	0.7	–	–
	Old image	33.8	44.5	21.0	0.7	4.8	17.1
	Old text	32.8	40.6	25.9	0.7	3.8	13.2
	Post-state text	51.2	7.7	40.4	0.7	22.2	−19.7
	No extra image	28.8	26.8	43.7	0.7	−0.2	−0.6
	Unrelated text	29.2	25.9	44.1	0.7	0.2	−1.5

Table E.6: **Candidate-by-evidence interaction (pp) in the same fixed-description experiment.** Each entry subtracts the correct-description accuracy effect from the unfinished-edit effect.

Model	Old image	Old text	Post-state text
Qwen3-VL-4B	31.0	22.7	24.0
Gemma-3-4B	19.2	14.1	14.9
Qwen3-VL-8B	22.9	16.8	17.8

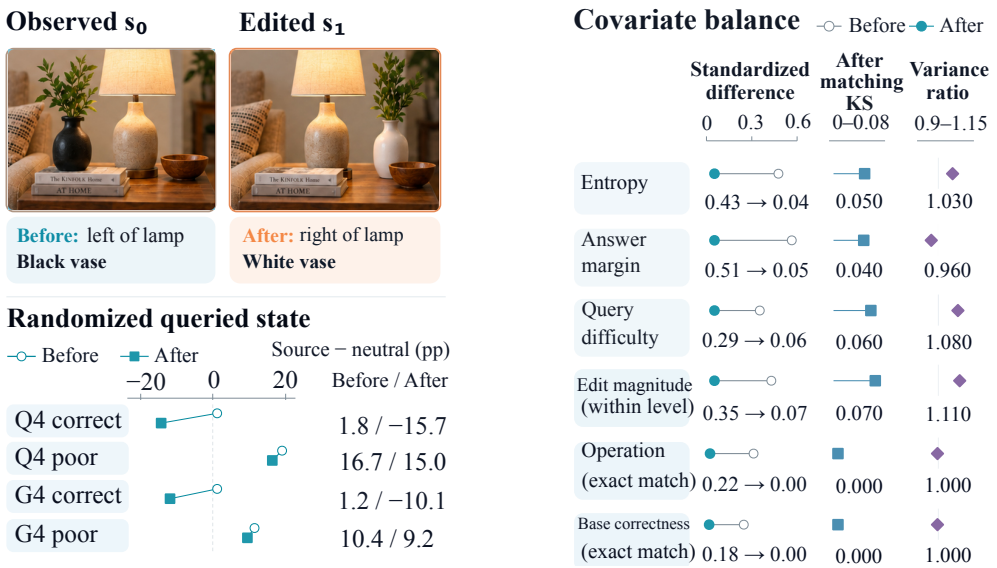
E.3 State-applicability controls

Table E.7 reports balance before and after matching. The post-match absolute standardized differences for entropy, answer margin, query difficulty, and edit magnitude are 0.04–0.07. Operation and base correctness are exactly matched. The post-match statistics quantify the remaining imbalance in each covariate.

Table E.7: **Covariate balance.** Absolute standardized differences, Kolmogorov–Smirnov statistics and variance ratios before and after matching.

Covariate	Standardized difference		After matching	
	Before	After	KS	Var. ratio
Entropy	0.43	0.04	0.05	1.03
Answer margin	0.51	0.05	0.04	0.96
Query difficulty	0.29	0.06	0.06	1.08
Edit magnitude within treatment level	0.35	0.07	0.07	1.11
Operation exact-match	0.22	0.00	0.00	1.00
Base correctness exact-match	0.18	0.00	0.00	1.00

The randomized before/after target comparison in Table 1 addresses a different question. Keeping candidate labels relative to the post-state, old evidence helps when the question asks about the before state and harms correct candidates when it asks about the after state. Figure E.2 illustrates how the



(a) Randomized target state

(b) Covariate balance

Figure E.2: **The requested state changes source value beyond uncertainty differences.** (a) Randomizing the queried state changes source-minus-neutral accuracy; correct and poor describe post-state candidates throughout. (b) Matching reduces covariate imbalance, with KS statistics and variance ratios reporting residual differences. Photographs illustrate the before/after semantics.

requested relation changes while the source remains truthful. Randomizing the target complements the approximate matching of pre-evidence covariates.

Table E.8 varies state labels, the strength of the rechecking instruction, evidence recency, message role, and matched padding. Explicit pre-edit labels retain the accuracy loss with correct descriptions and the gain with unfinished descriptions. The much smaller identity two-pass contrast and the negative random-residual control weaken explanations based on generic extra processing or perturbation magnitude alone.

Table E.8: **Evidence presentation and content controls.** Values are old-evidence-minus-neutral accuracy effects (pp); C/P denote correct/poor post-state candidates. Rows are distinct interventions.

Control	Qwen3-VL-4B		Gemma-3-4B	
	Correct	**Poor**	**Correct**	**Poor**
State-explicit labels	−16.0	15.0	−9.9	9.3
Weak check-again	−21.8	14.3	−13.9	8.6
Old first	−12.6	14.0	−7.6	8.4
Old last	−17.8	15.7	−11.2	9.7
Single message role	−14.9	14.8	−9.0	9.0
Masked equal padding	−15.8	15.1	−9.6	9.1
Complete analytic diagram	−13.2	13.8	−8.3	8.2
Complete coordinate table	−12.8	13.2	−7.7	8.0
Old blur sigma 1	−11.6	11.0	−7.6	6.7
Old blur sigma 2	−8.3	6.5	−5.4	4.3
Before construction	−5.3	10.8	−3.4	6.8
After candidate	−16.0	15.0	−9.9	9.3
Random norm-matched residual	−3.5	−1.8	−2.1	−1.1

Providing old evidence later increases the accuracy loss for correct post-state descriptions. Blur reduces both damage and repair, rather than selectively improving verification. This pattern is consistent with weakening a source that supplies both useful geometry and competing old relations.

Timing matters because using the source to execute an edit is different from letting it compete with an already constructed state.

Analytic diagrams and coordinate tables yield nearby, quality-dependent effects. Old text also changes the answer distribution in Tables E.3–E.5. The shared pattern supports state applicability across these evidence modalities. A natural photograph can contain facts omitted by a short text description, so image/text differences are not interpreted as a pure modality effect. Text and geometry therefore provide substantive alternatives in the capability comparisons.

E.4 MATCHED MARGINAL ERRORS

For controlled scenes, coordinates contain object centers and orientation vectors in the specified frame, normalized using training-set scales and equally weighted in W. Candidate replicas are formed without the terminal query. Within the same canonical scene and edit, pairing candidate and source-readout errors creates shared, independent, and opposed conditions. The source is not replaced by an unrelated scene. Candidate replicas remain nested within their source cluster.

Table E.9 reports the measured moments and low-dose risk changes. Qwen's shared and independent conditions have nearby A, D but cross-moments 1.72 and 0.03, with risk changes $+0.33$ and -0.19. The opposed condition further reduces risk. Gemma has the same qualitative ordering. These neural conditions approximately match moments rather than establishing identity of complete marginal distributions; Table A.1 gives the exact fixed-distribution construction.

Table E.9: **Matched-marginal neural intervention at source-key dose $\lambda = 0.25$.** A, D are marginal second moments and B is the cross-moment. Risk changes use the candidate as reference.

Family / coupling	Measured moments			Risk at $\lambda = 0.25$	
	A	D	B	R	ΔR
Q4 shared	1.01	4.02	1.72	1.34	0.33
Q4 independent	1.00	4.00	0.03	0.81	-0.19
Q4 opposed	1.02	4.04	-1.69	0.43	-0.59
G4 shared	1.03	3.98	1.58	1.32	0.29
G4 independent	1.01	4.03	0.06	0.84	-0.17
G4 opposed	1.00	4.01	-1.55	0.46	-0.54

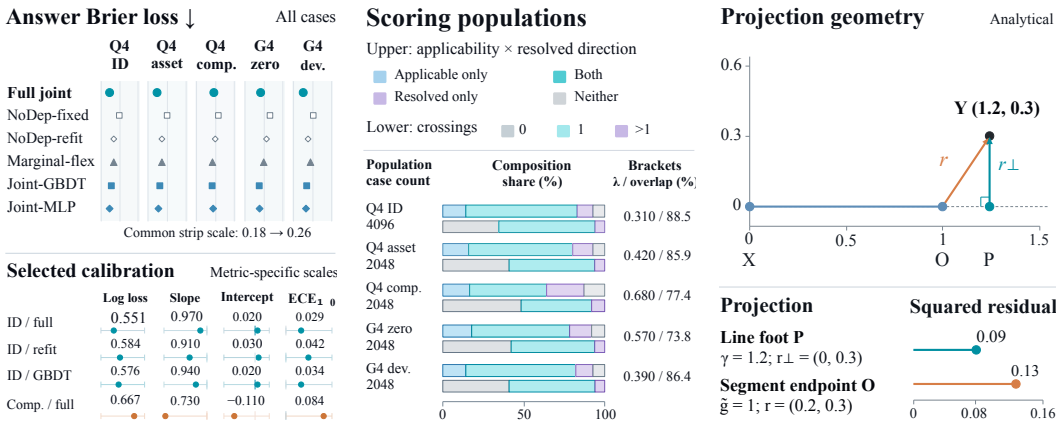
Figure 2 places the three complementary comparisons side by side. Its first panel fixes complete marginal distributions; its second reports measured moment-matched responses; its third separates risk and answer boundaries. The measured pairing intervention tests the sign contrast predicted by the exact construction. Table E.10 separates bias and covariance. Independence does not generally imply $B = 0$: its baseline is $B_{\text{ind}} = \mu_e^\top W \mu_z$. In the listed biased condition, the mean product is 0.03 and the independent cross-moment is 0.04. Thus a dependence ablation must preserve the means rather than silently centering away part of the original error.

Table E.10: **Marginal matching and biased errors.** Exact rows have identical finite marginals. Neural rows report KS distances, means, cross-moment, and covariance. Independence preserves the mean-product term rather than setting every cross-moment to zero.

Condition	Marginal KS		Mean error		Joint error	
	e	z	μ_e	μ_z	B	**Cov.**
Exact independent centered	0.000	0.000	0.000	0.000	0.000	0.000
Exact shared centered	0.000	0.000	0.000	0.000	2.000	2.000
Neural independent-like biased	0.028	0.024	0.100	0.300	0.040	0.010
Neural shared biased	0.035	0.031	0.100	0.300	1.720	1.690
Neural opposed biased	0.032	0.029	0.100	0.300	-1.690	-1.720

E.5 HELD-OUT ANSWER PROBABILITIES

The prediction instrument connects a geometric response to an answer probability. Development-only calibration specifies the response scale and joint margin distribution; held-out evaluation tests both. Figure E.3 separates prediction accuracy, scoring populations, and response residuals.



(a) Answer prediction (b) Scoring populations (c) Projection residuals

Figure E.3: **Prediction accuracy, scoring populations, and response residuals are distinct diagnostics.** (a) Joint and marginal predictors are evaluated by all-case Brier loss and calibration. (b) Applicability and resolved direction define different subsets; crossing counts partition all cases. (c) The analytical segment residual combines off-line displacement with movement beyond an endpoint, unlike the orthogonal line residual.

The three-parameter response map is fitted on prediction-development coordinate readouts,

$$g_\theta(\lambda, h) = \sigma\big(\theta_0 + \theta_1 \log(\lambda + 10^{-3}) + \theta_2 h\big)\,, \qquad \theta_1 \geq 0, \tag{E.1}$$

where h is the candidate-error level. The geometry needed to measure errors is available to every predictor in this controlled comparison. It is not an input to deployed counterfactual answering on photographs. Hyperparameter development for trained systems is separate from predictor development.

A depth-three conditional tree stratifies operation, conflict, and candidate-error level. Each leaf retains empirical joint residual pairs from at least 64 source clusters. Eight candidate/perception replicas per source characterize within-source variation; they are not independent training runs. The predictor uses 256 fixed numerical integration draws from the development residual library to propagate query margins. For a binary affine query with correct target margin $m_q(s_1)$, each propagated margin is

$$M_g = m_q(s_1) + (1 - g)w_q^\top E + g w_q^\top Z. \tag{E.2}$$

Equation (E.2) turns the response map in Eq. (E.1) into a query-specific event: the fraction of positive margins gives a raw correctness probability. For multiple choices, correctness requires all correct-versus-competitor margins to be positive simultaneously; their probabilities are not multiplied as though independent. Following development-set probability calibration (Guo et al., 2017), a two-parameter logistic map is fitted by source-disjoint cross-fitting. The resulting predictor uses the joint margin distribution in addition to the three risk moments.

The full predictor preserves joint pairs. The fixed independence control reshuffles the source residual pairing while freezing the tree, response map, and final calibration. The refitted control preserves the marginal distributions and their means but refits the same-capacity response map and calibration under the independent pairing. The flexible marginal comparator receives marginal information without pairwise features that reconstruct dependence. Joint GBDT and MLP comparators receive the same controlled geometric information. The GBDT uses depth three, 100 trees, and learning rate 0.05; the MLP uses two width-64 layers. Each learned comparator has six development configurations. Numerical integration draws are fixed before test outcomes and are not added to the test denominator.

Table E.11 reports coordinate-response error alongside the answer Brier comparison in Table 2. The full model improves over both independence variants on ID and asset shift. On composition and zero-calibration family transfer, flexible joint predictors are stronger. Recalibrating on Gemma development data restores a favorable comparison. Breaking error pairs loses information within a calibrated regime. Transferring the response map introduces a separate error even when those pairs are retained.

Table E.11: **Held-out coordinate-risk mean absolute error ↓.** The fixed independence intervention holds calibration constant; refitting gives the independence-restricted model a separate development fit.

Population	Structural response			Flexible predictors		
	Joint	**Indep. fixed**	**Indep. refit**	**Marginal**	**Joint GBDT**	**Joint MLP**
Qwen: ID	0.073	0.128	0.108	0.110	0.084	0.081
Qwen: asset shift	0.092	0.151	0.131	0.133	0.104	0.099
Qwen: composition	0.190	0.248	0.216	0.188	0.159	0.164
Gemma: zero calibration	0.162	0.225	0.201	0.154	0.124	0.130
Gemma: dev. calibration	0.081	0.139	0.116	0.119	0.096	0.094

Table E.12 reports log loss, calibration slope/intercept, and calibration error for the answer probabilities. The full model's composition slope of 0.73 and zero-calibration Gemma slope of 0.71 accompany weaker proper losses. Family calibration raises the slope to 0.95. The probability map needs recalibration even when coordinate responses retain their directional ordering.

Table E.12: **Answer-probability calibration.** Log loss and ten-bin expected calibration error (ECE) are lower-is-better; ideal calibration has slope one and intercept zero.

Population / predictor	**Log loss**	Probability calibration		
		Slope	**Intercept**	**ECE$_{10}$**
ID full	0.551	0.970	0.020	0.029
ID No-dep.-refit	0.584	0.910	0.030	0.042
ID joint-GBDT	0.576	0.940	0.020	0.034
Composition full	0.667	0.730	−0.110	0.084
Composition joint-GBDT	0.626	0.890	−0.040	0.045
Gemma zero full	0.651	0.710	−0.140	0.093
Gemma zero joint-GBDT	0.614	0.900	−0.020	0.043
Gemma calibrated full	0.569	0.950	0.020	0.033

E.6 Scoring populations and crossings

Table E.13 gives the denominator of each prediction task. All-case proper losses retain out-of-class, unresolved, and no-crossing cases. Applicability is defined by input/development criteria. Resolved direction instead requires an observed normalized risk change outside the ± 0.02 tie band. These events and their intersection are reported separately; neither defines an independent experimental unit.

Table E.14 scores direction only on the resolved population. For Qwen ID, that denominator is 3,359 rather than all 4,096 cases. The full predictor reaches 92.1% there, compared with 82.7% for fixed independence and 85.6% for refitted independence. Composition reverses its ordering against the joint GBDT. These conditional direction accuracies do not replace the all-case probability losses.

Table E.15 retains no crossing, a single crossing, and multiple crossings. The categories sum to the all-case denominator, but only a crossing supplies a finite-grid bracket. ID contains 1,475 no-crossing cases and 102 multiple-crossing cases. Composition has more multiple crossings and wider brackets. A bracket-overlap rate is not a confidence interval for a mean, and the finite dose grid does not reveal an exact continuous root.

Table E.13: **Prediction scoring populations.** Applicability uses input/development criteria; resolved direction requires measured risk change outside the tie band; Both is their intersection. All-case probability losses include every case, including those with no crossing.

Population	Case counts				
	All	**Applicable**	**Resolved**	**Both**	**No crossing**
Qwen: ID	4096	3604	3359	3028	1475
Qwen: asset shift	2048	1720	1622	1388	881
Qwen: composition	2048	1352	1434	996	983
Gemma: zero calibration	2048	1638	1556	1270	922
Gemma: dev. calibration	2048	1741	1659	1438	799

Table E.14: **Resolved-direction accuracy (%) ↑.** The tie band is ± 0.02 normalized risk units. Each row uses its resolved-direction denominator from Table E.13.

Population	Structural response			Flexible
	Joint	**Indep. fixed**	**Indep. refit**	**Joint GBDT**
Qwen: ID	92.1	82.7	85.6	84.8
Qwen: asset shift	89.4	78.8	82.1	82.5
Qwen: composition	78.2	69.4	73.1	79.5
Gemma: zero calibration	79.8	70.6	75.4	81.2
Gemma: dev. calibration	89.9	79.6	83.0	82.4

Table E.15: **Finite-dose crossing diagnostics.** No, single, and multiple crossings partition each population. Bracket widths use attention-dose units λ; overlap measures finite-grid agreement rather than continuous-root coverage.

Population	Observed crossing count			Finite-grid brackets	
	None	**Single**	**Multiple**	**Width (λ)**	**Overlap (%)**
Qwen: ID	1475	2519	102	0.31	88.5
Qwen: asset shift	881	1085	82	0.42	85.9
Qwen: composition	983	921	144	0.68	77.4
Gemma: zero calibration	922	1033	93	0.57	73.8
Gemma: dev. calibration	799	1171	78	0.39	86.4

E.7 Source dose and state-aware recomputation

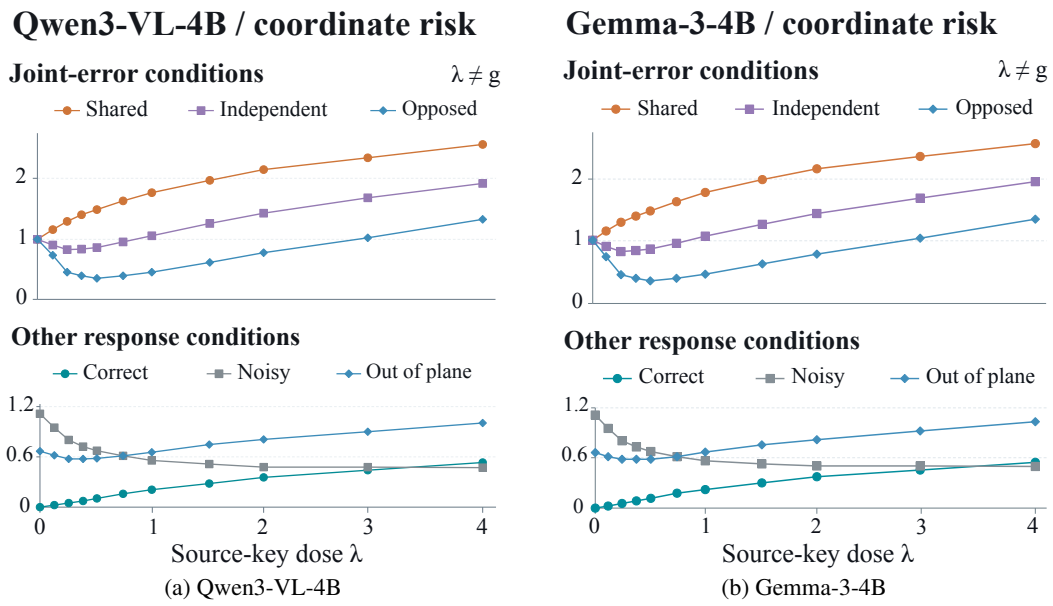
Table E.16 reports absolute coordinate risks at the six recorded source-key doses. Independent errors first improve and later lose the benefit, shared errors worsen from low dose, and opposed errors have a wider useful range. Each condition uses its own zero-dose readout as reference. Figure E.4 displays these profiles with linearly interpolated intermediate points. Interpolation is used for presentation, not as additional measurements or finer empirical crossing estimates.

The fixed-dose moment-matched intervention in Table E.9 tests error pairing; these dose profiles test how each response changes with source-key attention. Their aggregate endpoints are used within their respective comparisons. The calibrated prediction study reports its own scoring populations and finite-dose brackets in Appendix E.6.

Table E.17 distinguishes raw mixing, state transport, and recomputation. For the collinear under-edit, raw-state risk is 1.44, transported-state risk is 0.64, and recomputation gives 0.48. Applying the transformation recovers a direction unavailable to raw mixing, as specified by Lemma B.3. The out-of-line condition favors recomputation more strongly. These comparisons identify a state-aware repair mechanism beyond source-directed movement.

Table E.16: **Six-dose coordinate-risk study.** Headers denote source-key dose λ and entries are absolute risks. The profiles use six recorded nodes; intermediate display points add no measurements.

Condition	Source-key dose λ					
	0	**0.25**	**0.5**	**1**	**2**	**4**
Qwen3-VL-4B						
correct high conflict	0.000	0.031	0.075	0.162	0.302	0.484
noisy low conflict	1.100	0.762	0.633	0.515	0.440	0.419
shared error high	1.000	1.293	1.494	1.792	2.174	2.588
independent same norm	1.000	0.802	0.842	1.028	1.402	1.913
opposite error	1.000	0.403	0.310	0.395	0.737	1.292
out of plane	0.650	0.545	0.550	0.612	0.747	0.940
Gemma-3-4B						
correct high conflict	0.000	0.034	0.081	0.174	0.325	0.522
noisy low conflict	1.100	0.767	0.644	0.536	0.475	0.471
shared error high	1.000	1.295	1.498	1.801	2.190	2.614
independent same norm	1.000	0.805	0.850	1.043	1.425	1.947
opposite error	1.000	0.409	0.321	0.411	0.759	1.319
out of plane	0.650	0.549	0.559	0.625	0.766	0.965



(a) Qwen3-VL-4B (b) Gemma-3-4B

Figure E.4: **Increasing source dose can exhaust a low-dose benefit.** Profiles connect the six recorded dose nodes in Table E.16; intermediate markers are linear display interpolation, not additional measurements. Each condition uses its own zero-dose readout. Source-key attention dose and geometric response strength are distinct quantities.

Table E.17: **Raw-source, transported-source, and recomputation responses.** Risks are coordinate diagnostics; the last column is the answer-change rate induced by coordinate probing (percentage points), evaluated through a separate fresh branch.

Response condition	Coordinate risk			Probe change (pp)
	Raw	**Transport**	**Recompute**	
Collinear under-edit	1.44	0.64	0.48	0.8
Independent perception error	0.81	0.55	0.57	0.6
Correct high-conflict candidate	0.04	0.01	0.03	0.4
Out-of-line response	0.56	0.49	0.31	2.6

F CONSTRUCTION AND SOURCE RETENTION

Construction determines which post-state reaches the reader; retention determines whether the original state competes with it. Holding the construction and reader fixed isolates the second decision. The following controls then test whether secondary banks, textual history, training objectives, or parsing alter the information supplied at that interface.

F.1 FIXED-BANK SOURCE EXCHANGE

Tables F.1 and F.2 cross donor quality and conflict with frozen and mode-balanced readers. Balanced training presents equal shares of append and replacement on a source-disjoint split with the constructor frozen. Within a row, only the evidence-use interface changes. Blend coefficients give the original-bank fraction at the native interface; geometric response strength g refers to the separate coordinate model.

Table F.1: **Qwen3-VL-4B source retention with frozen and mode-balanced readers.** Accuracy (%) ↑ is compared at fixed donor banks and reader weights within each row. Bold/underline mark the two highest accuracy values per row; NMSE measures the donor, and blend coefficients specify the original-bank fraction.

Quality / conflict	Evidence interface			Direct latent blend			**NMSE**
	Replace	**Append**	**Padded**	**0.25**	**0.50**	**0.75**	
Frozen reader							
Poor / High	46.2	**64.1**	46.1	51.3	<u>53.5</u>	50.4	0.62
Poor / Low	46.2	**58.4**	46.1	51.3	<u>53.5</u>	50.4	0.62
Medium / High	68.4	64.8	<u>68.5</u>	**68.7**	67.0	65.3	0.24
Medium / Low	68.4	**73.0**	68.5	<u>68.7</u>	67.0	65.3	0.24
Good / High	**84.9**	63.6	<u>84.8</u>	79.6	73.8	66.2	0.06
Good / Low	<u>84.9</u>	**85.7**	84.8	84.5	83.8	82.3	0.06
Oracle / High	**93.1**	64.0	<u>93.0</u>	87.9	80.8	72.6	0.00
Oracle / Low	**93.1**	92.8	<u>93.0</u>	92.9	92.2	91.6	0.00
Mode-balanced reader							
Poor / High	47.0	**68.5**	46.9	52.1	<u>54.3</u>	51.2	0.62
Poor / Low	47.0	**61.1**	46.9	52.1	<u>54.3</u>	51.2	0.62
Medium / High	69.2	**71.1**	69.3	<u>69.5</u>	67.8	66.1	0.24
Medium / Low	69.2	**74.4**	69.3	<u>69.5</u>	67.8	66.1	0.24
Good / High	**85.6**	72.3	<u>85.5</u>	80.3	74.5	66.9	0.06
Good / Low	<u>85.6</u>	**86.1**	85.5	85.2	84.5	83.0	0.06
Oracle / High	**93.4**	73.5	<u>93.3</u>	88.2	81.1	72.9	0.00
Oracle / Low	**93.4**	93.0	<u>93.3</u>	93.2	92.5	91.9	0.00

With the frozen Qwen reader, high-conflict poor states benefit from append and good states from replacement. At the same good-state error, low conflict instead gives 85.7% with append and 84.9% with replacement. Balanced training narrows the good-state high-conflict gap, while preserving its direction. Padding-matched replacement remains close to ordinary replacement.

Table F.2: **Gemma-3-4B source retention with frozen and mode-balanced readers.** Accuracy (%) ↑ is compared at fixed donor banks and reader weights within each row. Bold/underline mark the two highest accuracy values per row; NMSE measures the donor, and blend coefficients specify the original-bank fraction.

Quality / conflict	Evidence interface			Direct latent blend			**NMSE**
	Replace	**Append**	**Padded**	**0.25**	**0.50**	**0.75**	
Frozen reader							
Poor / High	41.7	**58.9**	41.6	46.8	<u>49.0</u>	45.9	0.62
Poor / Low	41.7	**54.0**	41.6	46.8	<u>49.0</u>	45.9	0.62
Medium / High	64.0	61.6	<u>64.1</u>	**64.3**	62.6	60.9	0.24
Medium / Low	64.0	**68.4**	64.1	<u>64.3</u>	62.6	60.9	0.24
Good / High	**80.2**	59.8	<u>80.1</u>	74.9	69.1	61.5	0.06
Good / Low	<u>80.2</u>	**80.9**	80.1	79.8	79.1	77.6	0.06
Oracle / High	**89.4**	60.5	<u>89.3</u>	84.2	77.1	68.9	0.00
Oracle / Low	**89.4**	89.2	<u>89.3</u>	89.2	88.5	87.9	0.00
Mode-balanced reader							
Poor / High	43.0	**64.7**	42.9	48.1	<u>50.3</u>	47.2	0.62
Poor / Low	43.0	**57.6**	42.9	48.1	<u>50.3</u>	47.2	0.62
Medium / High	65.1	**67.9**	65.2	<u>65.4</u>	63.7	62.0	0.24
Medium / Low	65.1	**70.0**	65.2	<u>65.4</u>	63.7	62.0	0.24
Good / High	**81.1**	68.6	<u>81.0</u>	75.8	70.0	62.4	0.06
Good / Low	<u>81.1</u>	**82.0**	81.0	80.7	80.0	78.5	0.06
Oracle / High	**90.2**	70.7	<u>90.1</u>	85.0	77.9	69.7	0.00
Oracle / Low	**90.2**	89.9	<u>90.1</u>	90.0	89.3	88.7	0.00

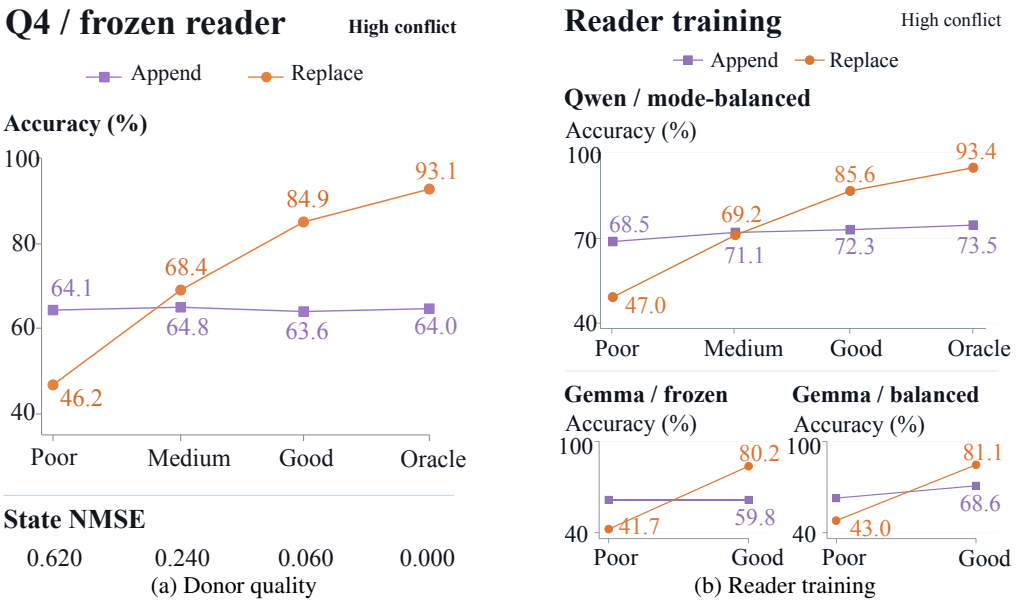
Gemma preserves the poor-state source benefit and reliable-state high-conflict replacement benefit under both reader settings. Oracle banks remove construction error yet retain a reading limitation. These comparisons separate donor quality, training interface, and the old relations available at inference; the numerical conditions, not the illustrative answer snippets, define the contrasts.

F.2 Quality-by-conflict interactions

Figure F.1 connects donor quality and reader training. Table F.3 varies normalized construction error more finely than the donor labels. At error 0.10, replacement exceeds append by 14.8 points under high conflict but trails it by 0.6 under low conflict. At error 0.60, both strata favor append. The useful boundary is consequently two-dimensional in these controls, rather than a universal quality threshold. The added reader conditions in Fig. F.1(b) ask whether the interaction is only a consequence of training for one interface. Mode-balanced training improves append on reliable states, but does not remove its high-conflict deficit. The oracle comparison asks the next question: can reading still fail after construction error has been removed?

Table F.3: **Construction error by edit conflict.** Replacement minus append accuracy (pp) for Qwen with a fixed donor and reader in each comparison.

State NMSE	Replacement minus append (pp)	
	High conflict	**Low conflict**
0.02	24.1	0.9
0.05	20.3	0.3
0.10	14.8	−0.6
0.18	8.2	−2.8
0.28	1.1	−4.9
0.40	−7.3	−7.2
0.60	−15.4	−10.2
0.80	−20.1	−11.3



(a) Donor quality (b) Reader training

Figure F.1: **Reader training changes the size, but not the existence, of the retention interaction.** (a) Each donor condition fixes banks and reader while exchanging append and replacement. (b) Mode-balanced training narrows the reliable-state gap. Oracle banks remove construction error while leaving the reader's evidence-use limitation visible.

The donor controls identify the relevant quality–conflict interaction. Appendix H.3 tests whether a public-input estimator can exploit it in a source-retention policy.

F.3 Old-state banks and textual history

Table F.4 isolates implementation routes. Replacing only the initial bank leaves a large deficit, and restoring the correct secondary banks recovers 57.5% on MindEdit and 71.9% on STARE, close to the complete system. Retaining stale caches also hurts. The recovery localizes the deficit to an unexchanged visual route.

Fresh textual computation tests another route for obsolete information. FSAF addresses stale reasoning when the observed image changes (Fan et al., 2026). Table F.5 crosses that setting with a hypothetical edit to an unchanged photograph. On observed image changes, the new-photo firewall reaches 88.0%, close to 88.4% for compatible banks plus the firewall. In the hypothetical setting, raw source plus the firewall reaches 77.2%, below 89.1% for compatible banks plus the same firewall. Removing stale text is therefore nearly sufficient in one comparison but leaves a state-relevance issue in the other.

Table F.4: **Scene-bank and cache ablations.** Complete-task Qwen accuracy (%) ↑ with the same public inputs and answer options.

Intervention	Complete-task accuracy (%) ↑	
	MindEdit-L4	**STARE-rigid**
Full all-bank fresh replacement	57.8	72.2
Initial bank only; stale secondary	49.1	65.4
All banks; stale KV retained fault	51.0	66.7
Padding matched replacement	57.7	72.3
Replace secondary with correct donor rescue	57.5	71.9
Unrelated same-norm secondary bank	46.8	61.9
Source banks with fresh text firewall	43.1	59.7

Table F.5: **Observed image change and hypothetical edit are different interventions.** This factorial control crosses visual evidence state with old text or a Fresh-State Attention Firewall (FSAF). Correct-card accuracy and old-answer share are separate marginals (%).

World update	Evidence / text policy	Answer share (%)	
		Correct	**Old a_0**
Hypothetical edit, same photo	Raw source+old text	68.7	21.6
	Raw source+FSAF	77.2	13.4
	Compatible banks+old text	81.3	9.5
	Compatible banks+FSAF	89.1	3.4
Observed image changes	New photo+old text	70.2	18.9
	New photo+FSAF	88.0	4.1
	Compatible banks+old text	73.8	15.7
	Compatible banks+FSAF	88.4	3.8

The two settings differ in whether the visible source already represents the target. The response to the same firewall therefore depends on which information is stale.

F.4 State supervision and warmup

Learning the target state may help beyond the warmup curriculum. Table F.6 crosses warmup choices with joint state supervision. This separates the joint objective from the curriculum, unlike simply removing a state loss and its warmup together. Under post-state warmup, adding joint state supervision raises Qwen from 51.7% to 57.8% on MindEdit and from 67.4% to 72.2% on STARE. Other warmup rows test whether that ordering depends on the initial phase.

Table F.6: **Warmup and joint state supervision.** Complete-task accuracy (%) ↑. The warmup objective applies to the first 128 updates; the binary indicator specifies state supervision during joint training.

Warmup objective	Joint state loss	Qwen3-VL-4B		Gemma-3-4B	
		L4	**STARE**	**L4**	**STARE**
Answer-only	0	48.7	64.8	43.3	59.8
Identity-state	0	49.6	65.0	44.1	60.0
Post-state	0	51.7	67.4	45.3	61.2
Answer-only	1	56.4	71.0	48.8	65.7
Identity-state	1	56.9	70.8	49.0	65.3
Post-state	1	57.8	72.2	50.2	66.5

Table F.7 holds the full configuration as a common reference for component changes. Uniform state weights give 57.5% on MindEdit and 72.4% on STARE, close to the full model. Both weighting rules retain the compatible-state benefit in these comparisons. Removing observation loss raises

counterfactual accuracy while reducing invariant accuracy from 86.3% to 83.7%. Shuffled and identity targets are weaker than true post-state targets, separating target semantics from the mere presence of an auxiliary loss.

Table F.7: **Training-objective ablations.** Qwen accuracy (%) ↑ on counterfactual tasks and unchanged facts. Removing observation supervision trades factual retention for edited-question accuracy.

Variant	Accuracy (%) ↑		
	MindEdit-L4	**STARE-rigid**	**Invariant**
Full model	57.8	72.2	86.3
Uniform state weights	57.5	72.4	86.1
No observation loss	58.4	72.8	83.7
No true-post answer loss	55.1	68.6	85.8
Frozen reader	52.4	68.1	86.6
Same-norm shuffled state targets	47.8	64.1	83.9
Identity targets throughout	49.2	65.4	86.5
No jointstate after post-state warmup	51.7	67.4	85.9

Table F.8 crosses donor construction, reader training, and use mode. It tests whether a gain follows the state, the reader, or their combination rather than treating an end-to-end score as a construction-only effect. Both construction and reader learning contribute, while retaining the source remains a separate decision after their weights are fixed.

Table F.8: **Constructor–reader transfer.** Each donor checkpoint is reused with readers trained for append, replacement, both modes, or neither. Test columns change use while holding that donor–reader pair fixed (accuracy, %).

Family	Donor training	Reader training	Accuracy (%) ↑	
			Replace	**Append**
Qwen3-VL-4B	Append	Append	52.9	46.0
	Append	Replace	54.2	43.8
	Append	Balanced	53.7	48.0
	Append	Frozen	47.6	42.9
	Replace	Append	55.0	46.8
	Replace	Replace	57.8	44.7
	Replace	Balanced	56.5	48.2
	Replace	Frozen	51.1	41.0
Gemma-3-4B	Append	Append	45.8	39.9
	Append	Replace	47.1	38.1
	Append	Balanced	46.9	41.7
	Append	Frozen	40.2	36.4
	Replace	Append	48.4	40.3
	Replace	Replace	50.2	38.7
	Replace	Balanced	49.1	42.6
	Replace	Frozen	44.5	35.9

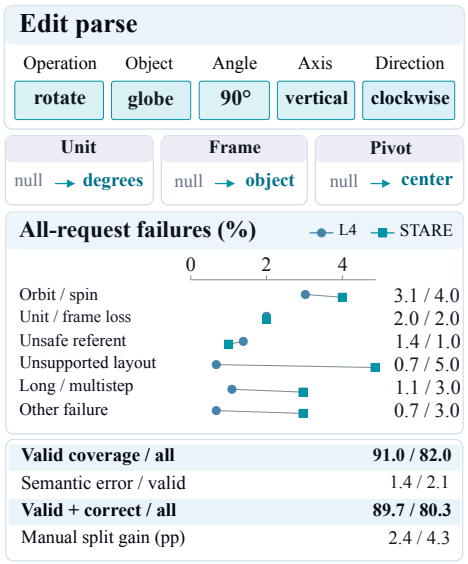
F.5 Parsing and visual controls

Figure F.2 links edit parsing with visual-input controls. Table F.9 partitions unsupported requests and separates syntactic coverage from semantic correctness. Orbit/spin confusion, a lost unit or frame, unsafe referents, unsupported layouts, and long edit chains have different consequences. A valid parse can still be wrong. Manual splitting improves the full score by 2.4 and 4.3 points, but supplies an oracle interface and cannot replace the automatic endpoint.

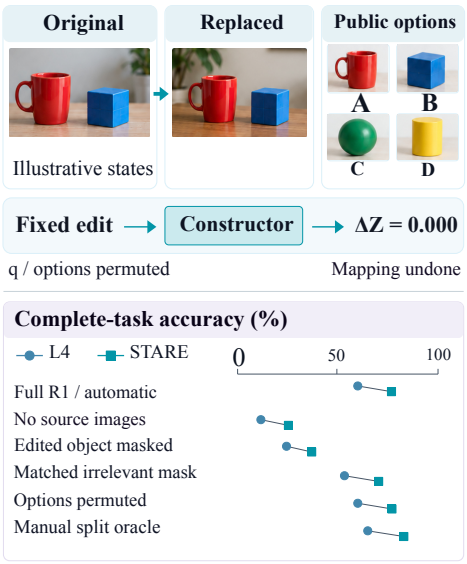
Table F.9: **Parser outcomes and full-task coverage.** Rejection categories and valid coverage use all requests. Semantic-error rates condition on valid parses; manual splitting is a diagnostic.

Metric	MindEdit-L4	STARE-rigid
Orbit/spin ambiguity	3.1	4.0
Lost unit/frame	2.0	2.0
Unsafe referent	1.4	1.0
Unsupported layout	0.7	5.0
Long/multistep	1.1	3.0
Other parse failure	0.7	3.0
Automatic valid coverage	91.0	82.0
Semantic error within valid parses	1.4	2.1
Valid and semantically correct (all)	89.7	80.3
Manual text split diagnostic gain	2.4	4.3

Table F.10 tests whether the model actually needs the scene and edit. Removing scene images while retaining option images greatly lowers accuracy; masking the edited object hurts more than masking an irrelevant object. Wrong-edit instructions also lower performance. Option permutation preserves the option content and largely preserves accuracy, weakening a fixed-position shortcut explanation. Figure F.2 distinguishes literal instruction parsing from visual necessity. A syntactically valid edit can still lose a pivot or frame, while an accurately parsed edit can still be answered without using the relevant object. Parsing determines the transformation that reaches the constructor; masking tests whether the reader then uses the object that transformation concerns.



(a) Parsing and coverage



(b) Input separation

Figure F.2: **A valid state interface must preserve edits while excluding answers from construction.** (a) Parsing checks units, frames, pivots, and referents; coverage is unconditional and semantic-error rates condition on valid parses. (b) Scene masking tests visual necessity. Fixed-edit query/option permutations test constructor agreement separately from answer accuracy.

Table F.10: **Visual necessity and interface controls for Qwen3-VL-4B.** Entries are accuracy (%) ↑ except the final row, which is an exact constructor-output mismatch norm. Option images remain legal public inputs.

Input or control	MindEdit-L4	STARE-rigid
Full automatic system	57.8	72.2
No scene images; keep options	12.6	23.8
Edited object masked	25.7	38.4
Matched area irrelevant object masked	54.8	70.0
Option permutation, mapping undone	57.1	71.8
Query paraphrase through parser	55.9	70.8
Manual split oracle	60.2	76.5
Wrong edit instruction control	18.4	30.1
Fixed-edit constructor mismatch norm	0.0	0.0

The zero constructor-output mismatch under a fixed-edit query permutation follows from excluding the terminal query from construction. The preceding parser still reads the original question. Task sensitivity to paraphrases and the benefit of manual splitting therefore locate a separate source of variation before construction.

F.6 Unedited capability

Table F.11 evaluates the trained reader on BLINK, ConVBench, and controlled invariant relations (Fu et al.; Jing et al., 2026). Replacement remains near the base score, whereas removing observation loss consistently weakens retention. On Qwen invariant relations, the full model gives 86.3% and the no-observation-loss variant gives 83.7%, despite the latter's slightly better counterfactual endpoints. Observation supervision retains factual uses that would otherwise be weakened by counterfactual training.

Table F.11: **Unedited-capability retention (%).** BLINK and invariant relations use accuracy; ConVBench uses consistency. Evaluation uses the same trained reader. Entries are point estimates, without a non-inferiority test.

Model	Endpoint	Accuracy (%) ↑				
		Base	**SFT**	**Append**	**Replace**	**No obs. loss**
Qwen3-VL-4B	BLINK accuracy	57.2	57.0	56.7	56.9	54.2
	ConVBench consistency	63.1	63.0	62.7	63.0	60.9
	CONTROL invariant accuracy	86.4	86.2	85.9	86.3	83.7
Gemma-3-4B	BLINK accuracy	52.4	52.3	51.9	52.0	49.9
	ConVBench consistency	59.5	59.3	59.0	59.4	57.3
	CONTROL invariant accuracy	82.0	81.9	81.5	81.8	79.4
InternVL3.5-8B	CONTROL invariant accuracy	87.1	86.8	86.4	86.7	84.0

The comparisons use the trained reader even on unedited tasks. Returning to the original pretrained checkpoint would test a different system. The listed tasks probe factual retention alongside the counterfactual improvement.

G COMPLETE-TASK CAPABILITY

Table 3 measures the capability of complete trained systems. The following comparisons retain unsupported requests, expand to other model families and task populations, and then examine external resources and representation choices. The question is where a compatible-state interface remains useful once construction, reading, and fallback all contribute to the answer.

G.1 CONSTRUCTION, READING, AND FALLBACK

Tables G.1 and G.2 report both tasks for Qwen and Gemma. Full scores include every request; parsed and fallback columns condition on the parser outcome. Repair and damage compare each separately trained method with paired SFT on the full request set. The transition partition describes a trained-system difference, whereas the fixed-bank experiments isolate source availability.

Table G.1: **Qwen3-VL-4B complete-task outcomes.** Full/Parsed/Fallback are accuracy (%); Full includes every request. Repair/Damage are paired-SFT transition shares (%), and Net is accuracy change (pp). P uses paired data; F, G, and E add filtering experts, geometry labels, and a geometry expert. Bold/underline rank Full within each P block.

Method	Accuracy by input path (%)			Transitions vs. SFT		
	Full	**Parsed**	**Fallback**	**Repair**	**Damage**	**Net**
MindEdit-L4						
Paired image/answer supervision (P)						
Paired SFT	33.1	33.9	25.0	0.0	0.0	0.0
Labeled append	46.0	48.1	24.8	14.8	1.9	12.9
Capacity, answer only	48.7	51.0	25.3	18.2	2.6	15.6
Outcome GRPO	52.2	54.9	25.4	20.5	1.4	19.1
Mirage port (2026)	<u>54.9</u>	57.9	24.9	23.1	1.3	21.8
CVPD port (2026)	50.1	52.6	25.1	17.9	0.9	17.0
Predicted text, paired data	52.6	55.3	25.5	20.7	1.2	19.5
Post-state replacement (ours)	**57.8**	61.1	24.4	25.9	1.2	24.7
Additional labels or experts (P+G, P+F, E)						
Text with geometry labels	57.0	60.1	25.7	24.8	0.9	23.9
PVST port (2026)	49.8	52.2	25.4	17.5	0.8	16.7
Predicted-geometry executor	41.7	43.3	25.8	10.7	2.1	8.6
STARE-rigid-text						
Paired image/answer supervision (P)						
Paired SFT	51.7	55.6	34.0	0.0	0.0	0.0
Labeled append	62.0	68.2	33.8	12.8	2.5	10.3
Capacity, answer only	64.8	71.5	34.3	16.3	3.2	13.1
Outcome GRPO	66.3	73.3	34.4	16.6	2.0	14.6
Mirage port (2026)	<u>70.6</u>	78.7	33.9	20.8	1.9	18.9
CVPD port (2026)	66.5	73.6	34.1	16.3	1.5	14.8
Predicted text, paired data	68.9	76.5	34.5	19.0	1.8	17.2
Post-state replacement (ours)	**72.2**	80.7	33.4	22.3	1.8	20.5
Additional labels or experts (P+G, P+F, E)						
Text with geometry labels	75.4	84.3	34.7	25.2	1.5	23.7
PVST port (2026)	66.8	73.9	34.4	16.5	1.4	15.1
Predicted-geometry executor	70.2	78.0	34.8	21.2	2.7	18.5

Qwen replacement reaches 57.8% on MindEdit and 72.2% on STARE-rigid-text. It improves over append and paired SFT through both more repair and less damage. Mirage is the strongest paired-data port in these comparisons. Geometry-supervised text and the geometry expert have additional spatial inputs and are grouped separately.

Table G.2: **Gemma-3-4B complete-task outcomes.** Full/Parsed/Fallback are accuracy (%); Full includes every request. Repair/Damage are paired-SFT transition shares (%), and Net is accuracy change (pp). P uses paired data; F, G, and E add filtering experts, geometry labels, and a geometry expert. Bold/underline rank Full within each P block.

Method	Accuracy by input path (%)			Transitions vs. SFT		
	Full	**Parsed**	**Fallback**	**Repair**	**Damage**	**Net**
MindEdit-L4						
Paired image/answer supervision (P)						
Paired SFT	28.7	29.5	21.0	0.0	0.0	0.0
Labeled append	39.9	41.8	20.8	13.1	1.9	11.2
Capacity, answer only	43.3	45.5	21.3	17.2	2.6	14.6
Outcome GRPO	45.8	48.2	21.4	18.5	1.4	17.1
Mirage port (2026)	<u>47.7</u>	50.4	20.9	20.3	1.3	19.0
CVPD port (2026)	44.5	46.8	21.1	16.7	0.9	15.8
Predicted text, paired data	46.3	48.8	21.5	18.8	1.2	17.6
Post-state replacement (ours)	**50.2**	53.1	20.4	22.7	1.2	21.5
Additional labels or experts (P+G, P+F, E)						
Text with geometry labels	51.0	53.9	21.7	23.2	0.9	22.3
PVST port (2026)	44.2	46.5	21.4	16.3	0.8	15.5
Predicted-geometry executor	38.9	40.6	21.8	12.3	2.1	10.2
STARE-rigid-text						
Paired image/answer supervision (P)						
Paired SFT	47.0	50.5	31.0	0.0	0.0	0.0
Labeled append	56.8	62.5	30.8	12.3	2.5	9.8
Capacity, answer only	59.8	66.1	31.3	16.0	3.2	12.8
Outcome GRPO	61.5	68.1	31.4	16.5	2.0	14.5
Mirage port (2026)	<u>63.9</u>	71.1	30.9	18.8	1.9	16.9
CVPD port (2026)	62.2	69.0	31.1	16.7	1.5	15.2
Predicted text, paired data	63.2	70.2	31.5	18.0	1.8	16.2
Post-state replacement (ours)	**66.5**	74.4	30.4	21.3	1.8	19.5
Additional labels or experts (P+G, P+F, E)						
Text with geometry labels	70.0	78.4	31.7	24.5	1.5	23.0
PVST port (2026)	61.6	68.2	31.4	16.0	1.4	14.6
Predicted-geometry executor	67.7	75.6	31.8	23.4	2.7	20.7

Gemma replacement gains 10.3 points over append on MindEdit and 9.7 on STARE-rigid-text. The same full/parsed/fallback distinction is retained across backbones. Fallback uses the trained reader with the original images and question, rather than substituting a different pretrained checkpoint.

G.2 Larger backbones and a third family

Table G.3 evaluates Qwen3-VL-8B, InternVL3.5-8B, and the 7B Spatial-SSRL lineage (Bai et al., 2025; Wang et al., 2025b; Liu et al., 2026). Qwen replacement reaches 65.8% on MindEdit and 76.6% on STARE, with gains over append in both tasks. The third-family result extends the comparison beyond the two principal readers. Spatial-SSRL leaves only a small additional gain on STARE, agreeing with the strong-base boundary rather than contradicting the larger-backbone result.

Table G.4 gives the same-data Mirage ports on the larger and third-family backbones. These ports extend the latent-visual-token comparison without transferring scores from a different model family or benchmark. The corresponding replacement endpoints remain in Table G.3.

Table G.3: **Complete trained comparisons across larger and additional model families (accuracy, %).** Bold/underline rank within each backbone. Spatial-SSRL here is the 7B Qwen2.5-VL lineage, distinct from the native-verification reference.

Model	Method	Accuracy (%) ↑	
		L4	**STARE**
Qwen3-VL-8B	SFT	42.1	57.4
	Append	<u>55.0</u>	<u>67.1</u>
	Replace	**65.8**	**76.6**
InternVL3.5-8B	SFT	40.3	55.1
	Append	<u>50.6</u>	<u>62.6</u>
	Replace	**58.7**	**71.8**
Spatial-SSRL7B/Qwen2.5VL	SFT	44.6	61.7
	Append	<u>60.6</u>	<u>74.3</u>
	Replace	**62.1**	**74.9**

Table G.4: **Task adaptations on larger Qwen and third-family backbones (accuracy, %).** Methods are compared within each backbone; pretraining differs across model families.

Model	Task port	Accuracy (%) ↑	
		L4	**STARE**
Qwen3-VL-8B	Mirage port	63.1	74.8
InternVL3.5-8B	Mirage port	57.4	70.5

Table G.5 returns to fixed use with a mode-balanced InternVL reader. Poor construction favors append, while a good state under high conflict favors replacement; low conflict reverses the latter ordering. This reproduces the construction–conflict interaction at a fixed interface in the third family.

Table G.5: **Fixed-use comparison for the mode-balanced InternVL3.5-8B reader (accuracy, %).** Conflict and quality are distinct factors; bold marks the higher fixed-use accuracy per row. Oracle construction is diagnostic.

Construction / conflict	Accuracy (%) ↑		NMSE
	Append	**Replace**	
InternVL3.5 poor/high-conflict	**61.3**	44.8	0.59
InternVL3.5 good/high-conflict	66.1	**82.7**	0.07
InternVL3.5 good/low-conflict	**83.4**	82.7	0.07
InternVL3.5 oracle/high-conflict	71.6	**92.0**	0.00

G.3 Operation and benchmark scope

Table G.6 balances the operation classes to distinguish translation, pivot orbit, and rigid rotation. This diagnostic has its own weighting and is not a reconstruction of native benchmark frequencies. Replacement's MindEdit orbit result exceeds its translation result, whereas other methods have different operation profiles. The benefit is consequently not reducible to a general preference for one easy edit class.

Tables G.7 and G.8 retain the full-benchmark checkpoints. These checkpoints are distinct from the final trained comparison in Table 3. MindEdit observational and reference-frame tasks change little, and several non-rigid STARE tasks have small decreases. Task accuracy and F1 remain separate. The broad evaluation locates the gains primarily in rigid counterfactual reasoning.

Table G.8 gives the corresponding Gemma checkpoints. As for Qwen, the observational and counterfactual tasks retain their native metrics and are evaluated within the same checkpoint.

Table G.6: **Operation-balanced Qwen3-VL-4B accuracy (%) ↑.** Each domain gives its two operation classes equal weight, independently of native benchmark frequencies. Bold/underline rank the two highest accuracy point estimates within P; additional-resource rows are unranked.

Method	MindEdit-L4		STARE-rigid		Macro	
	Trans.	**Orbit**	**Trans.**	**Rotate**	**L4**	**STARE**
Paired image/answer supervision (P)						
Paired SFT	37.9	28.3	55.1	48.3	33.1	51.7
Labeled append	48.0	44.0	63.9	60.1	46.0	62.0
Capacity, answer only	50.0	47.4	66.2	63.4	48.7	64.8
Outcome GRPO	53.1	51.3	68.5	64.1	52.2	66.3
Mirage port (2026)	<u>54.8</u>	<u>55.0</u>	<u>71.7</u>	<u>69.5</u>	<u>54.9</u>	<u>70.6</u>
CVPD port (2026)	53.6	46.6	69.1	63.9	50.1	66.5
Predicted text, paired data	52.3	52.9	69.8	68.0	52.6	68.9
Post-state replacement (ours)	**56.6**	**59.0**	**72.8**	**71.6**	**57.8**	**72.2**
Additional labels or experts (P+G, P+F, E)						
Text with geometry labels	56.6	57.4	75.9	74.9	57.0	75.4
PVST port (2026)	53.7	45.9	68.9	64.7	49.8	66.8
Predicted-geometry executor	43.8	39.6	74.5	65.9	41.7	70.2

Table G.7: **Full-benchmark Qwen3-VL-4B checkpoint (%), distinct from the primary trained system.** MindEdit pooled accuracy includes its object-edit task; STARE retains task-specific F1 and accuracy.

Task	Metric	Full-benchmark checkpoint		
		SFT	**Append**	**Replace**
		MindEdit		
L1	Accuracy	38.5	38.2	37.9
L2	Accuracy	22.3	22.5	22.4
L3a	Accuracy	17.8	18.6	19.7
L3b	Accuracy	11.6	12.1	12.7
L5	Accuracy	17.1	16.9	16.6
Pooled	Accuracy	22.2	24.6	26.8
		STARE		
2D	Accuracy	42.0	42.2	41.7
CubeNet	F1	46.0	46.2	45.7
Tangram	F1	51.0	51.2	50.7
Temporal	Accuracy	37.0	37.2	36.7
Perspective	Accuracy	31.0	31.2	30.7
3D	Accuracy	38.4	42.6	46.2

Table G.8: **Full-benchmark Gemma-3-4B checkpoint (%), distinct from the primary trained system.** MindEdit pooled accuracy includes its object-edit task; STARE retains task-specific F1 and accuracy.

Task	Metric	Full-benchmark checkpoint		
		SFT	**Append**	**Replace**
		MindEdit		
L1	Accuracy	32.3	32.0	31.7
L2	Accuracy	18.5	18.7	18.6
L3a	Accuracy	14.7	15.5	16.6
L3b	Accuracy	9.2	9.7	10.3
L5	Accuracy	13.4	13.2	12.9
Pooled	Accuracy	18.6	20.4	22.4
		STARE		
2D	Accuracy	38.5	38.7	38.2
CubeNet	F1	43.0	43.2	42.7
Tangram	F1	49.0	49.2	48.7
Temporal	Accuracy	35.4	35.6	35.1
Perspective	Accuracy	29.0	29.2	28.7
3D	Accuracy	35.2	38.4	42.0

The pooled MindEdit score uses all task endpoints from the same full-benchmark checkpoint.

G.4 Independent real-photo evaluation

Table G.9 evaluates a photo cohort independent of MindEdit. Replacement reaches 57.1%, compared with 47.4% for append and 36.8% for paired SFT. Geometry-supervised text reaches 59.2%, and the Mirage port reaches 55.8%. The append–replacement gap persists on a second photo population, while geometry-supervised text remains higher.

Table G.9: **Independent real-photo evaluation (complete-task accuracy, %).** The cohort is separate from MindEdit. Text with geometry labels has additional supervision; the Mirage row is a same-data task port. Bold/underline rank the paired-data block only.

Method	**Accuracy (%) ↑**
Paired images and answers (P)	
Paired SFT	36.8
Append	47.4
Mirage port	<u>55.8</u>
Replacement	**57.1**
Additional labels or experts	
Text with geometry labels	59.2
Predicted-geometry executor	48.6

Table G.10 checks ordinary verification on that cohort. Qwen's 4.8-point gain combines 7.2% repair and 2.4% damage, including 1.5% confirmed old-answer return. Gemma likewise has both a positive net effect and directed rollback. Source verification thus retains directed rollback on this cohort as well as its positive average effect.

Table G.10: **Ordinary native verification on the independent real-photo cohort.** Before/after accuracy and transition shares are percentages of all requests. Confirmed-old, other, and unknown-old errors partition damage.

Model	Accuracy (%) ↑		Transitions (%)		Damage partition (%)		
	Before	**After**	**Repair**	**Damage**	**Old a_0**	**Other**	**Unk.**
Qwen3-VL-4B	24.6	29.4	7.2	2.4	1.5	0.6	0.3
Gemma3-4B	20.8	24.7	5.9	2.0	0.9	0.6	0.5

G.5 Peer-reviewed systems and information

Table G.11 asks what changes when a comparator acquires another source of information or performs extra test-time work. Its resource column is part of the comparison, not an implementation footnote. ReRe's three-photo adaptation obtains predicted geometry through VGGT from the permitted photographs (Ma et al., 2026; Wang et al., 2025a). Lookback allocates further visual search (Bi et al., 2026), while MindJourney uses generated world-model views (Yang et al., 2025). A complementary view can recover the original scene more accurately; answering the counterfactual still requires applying the hypothetical edit.

Table G.11: **Peer-reviewed system adaptations (accuracy, %).** Scores are measured on the present tasks. Additional pretraining, experts, view synthesis, and test-time work are specified by row and define separate information conditions.

System / adaptation	Additional information or work	Accuracy (%) ↑	
		L4	**STARE**
ReRe three-photo adapter	VGGT; three public photographs	33.6	53.8
Lookback adaptive search	Adaptive search; all branches counted	38.2	56.3
MindJourney scope adapter	World model; no extra real views	35.5	57.6
IPT answer-only task adaptation	BAGEL training; no image decoding	45.7	66.8
IPT generated-image task adaptation	BAGEL and image decoding	47.3	67.9
ReasonBrain same-reader adapter	FLUX and expert supervision	48.4	66.5
3ViewSense canonical-view adapter	Predicted canonical views; teacher	40.5	73.8
3DThinker final-stage adapter	VGGT teacher; final-stage model	39.6	60.4
Spatial-SSRL7B native	Spatial RGB-D post-training	35.1	53.7
SymPL scope adapter	Predicted-geometry expert	42.2	73.6
TokenWarp scope port	Predicted depth; task implementation	39.4	66.4

The image-generation rows instead provide a transformed representation to the common reader. ReasonBrain uses reasoning supervision and an image generator (He et al., 2026). The IPT comparison separates imaginative supervision from test-time image decoding: its answer-only and generated-image adaptations differ in whether the reader receives a decoded image (Bigverdi et al., 2025; 2026). This is the relevant comparison for a state-compatible interface, since constructing a useful state need not require rendering pixels.

The geometry-oriented rows expose a different information advantage. Canonical views, symbolic layouts, and depth-supported token warping provide explicit spatial structure (Zhan et al., 2026; Jang et al., 2026; Lee et al., 2026). 3DThinker additionally uses a geometry teacher (Chen et al., 2026). Each adaptation retains its declared inputs and operation semantics; a camera-frame change does not implement an object orbit. Spatial-SSRL adds spatial post-training (Liu et al., 2026). Their endpoints consequently answer different resource-permission questions from the paired-data block in Table 3.

The fixed-card factorial in Table F.5 crosses textual-history isolation with compatible banks (Fan et al., 2026). This is an intervention on two sources of stale information, separate from the published-system comparison. Self-refinement and verification depend on the feedback available to the reader (Madaan et al., 2023; Huang et al., 2024; Gou et al., 2024; Wu et al., 2024). The factorial tests the specific remaining issue here: a freshly processed photograph can still describe the pre-edit world.

Table G.12 contrasts a smaller learned adapter with an additional geometric route. ScAle rescales activations in a frozen backbone (Chowdhury et al., 2026); geometry refinement adds a predicted-geometry expert. The comparison separates adapting the existing reader from supplying more explicit spatial information, without introducing an oracle post-state.

Table G.12: **Minimal-adapter and geometry-refinement comparisons (accuracy, %).** Scalar adaptation modifies the existing reader; the geometry expert supplies an additional predicted representation. Neither receives an oracle post-state.

Comparator	Permissions	Accuracy (%) ↑	
		L4	**STARE**
ScAle task port Q4	P; frozen backbone, scalar adapter	38.9	58.7
ScAle task port G4	P; frozen backbone, scalar adapter	34.4	53.2
Geometry verifier-refiner	Predicted-geometry expert	44.8	73.8

G.6 Evidence representations

Table G.13 compares source plus composed query, paired-data text, latent replacement, geometry-supervised representations, generated pixels, and explicit execution. Query composition $q \circ T_u$ is already a state-compatible route (Reiter, 2001). The question is which representation preserves the information needed for the task, not which one is most similar to an image.

Table G.13: **Representation and information permissions on Qwen3-VL-4B (accuracy, %).** Oracle text and banks receive the true target state and are excluded from deployment rankings. Geometric labels are an additional training resource. Bold and underlining rank accuracy ↑ within the paired-data block only.

Representation	Information	Accuracy (%) ↑	
		L4	**STARE**
Paired image/answer supervision (P)			
Source with $q \circ T_u$	P; no extra expert	49.3	67.6
Predicted text	P only	52.6	68.9
Predicted object tokens	P; same predicted extractor	<u>53.5</u>	<u>69.8</u>
Latent replacement (ours)	P	**57.8**	**72.2**
Additional labels, pretraining, or experts			
Text + geometric labels	P+G	57.0	75.4
Latent + same geometric labels	P+G	59.2	75.0
Pixel-state same-reader	P+G; generative pretraining	55.6	71.9
Predicted-geometry executor	E	41.7	70.2
Oracle target-state inputs			
GT post-state text	Oracle, evaluation only	89.4	94.2
GT post-state banks	Oracle, evaluation only	92.0	96.1

Latent replacement exceeds paired-data text in this comparison. Adding geometry labels changes the ordering and improves both text and latent variants, so representation and supervision must not be conflated. Oracle text and banks have much higher accuracy but supply the target state at test time; they diagnose construction and use limits rather than enter deployment rankings. Appendix H.3 separately evaluates source selection. The geometry-supervised latent arm changes training information as well as representation.

H TRANSFER AND SEQUENTIAL EDITS

An edited state must remain usable when the input appearance changes, when another edit follows, and when a policy chooses how to use its source. Appearance shifts test perception and construction. Sequential edits test the information retained in intermediate states, while routing tests whether public inputs identify the useful action on the intended population.

H.1 APPEARANCE AND OPERATION SHIFTS

Figure H.1(a) illustrates the six operation semantics; panel (b) compares distribution shifts above the predicted/oracle composition diagnostic analyzed in Section H.2. Replacement retains an advantage under asset, camera, and background shifts, but its margin contracts under composition. The unseen-operation condition reverses its ordering against append in both families. These failures would be obscured by a single pooled out-of-distribution score.

Tables H.1 and H.2 retain every operation endpoint. Under composition, Qwen replacement improves the macro score from 58.0% to 60.7%, yet lowers orbit accuracy from 46.4% to 42.8%. Gemma has the analogous orbit loss. Removal improves instead. The macro gain is therefore not evidence that every edited relation transfers.

Table H.1: **Qwen3-VL-4B accuracy under distribution shift (%).** Each shift equally weights six operation classes; bold/underline rank within that shift. Appearance, camera, composition, and unseen-operation shifts are evaluated separately.

Shift	Method	Translations		Object edits			View	Macro
		Rel.	Cam.	Orbit	Spin	Remove		
ID	SFT	73.2	68.2	62.2	75.2	79.2	63.2	70.2
	Append	78.9	72.7	68.3	79.4	83.8	67.5	75.1
	Replace	**85.6**	**82.3**	**77.5**	**88.2**	**91.8**	**76.2**	**83.6**
Asset	SFT	68.1	61.1	55.1	70.1	75.1	55.1	64.1
	Append	72.3	64.7	58.2	74.6	78.6	58.4	67.8
	Replace	**78.5**	**71.0**	**63.8**	**82.0**	**88.1**	**64.8**	**74.7**
Camera	SFT	65.3	56.3	49.3	68.3	74.3	48.3	60.3
	Append	71.1	59.8	53.4	73.8	80.2	52.9	65.2
	Replace	**75.7**	**65.0**	**56.8**	**79.5**	**86.2**	**57.4**	**70.1**
Background	SFT	71.8	61.8	54.8	75.8	82.8	53.8	66.8
	Append	77.7	66.7	59.3	82.2	89.7	58.8	72.4
	Replace	**86.3**	**75.1**	**66.7**	**89.8**	**98.9**	**66.2**	**80.5**
Composition	SFT	57.1	47.1	39.1	60.1	68.1	41.1	52.1
	Append	64.0	52.4	**46.4**	64.9	74.2	46.1	58.0
	Replace	**66.4**	**56.1**	42.8	**70.3**	**80.5**	**48.1**	**60.7**
Unseen operation	SFT	52.9	45.9	38.9	54.9	60.9	39.9	48.9
	Append	**56.1**	**47.6**	**41.2**	**58.4**	64.5	**42.4**	**51.7**
	Replace	56.0	47.0	37.1	57.9	**65.9**	40.9	50.8

Camera shifts alter the geometric information available to perception, while background shifts primarily change appearance. These shifts stress different parts of the perception–construction pipeline, beyond the source-retention decision.

Table H.2: **Gemma-3-4B accuracy under distribution shift (%).** Each shift equally weights six operation classes; bold/underline rank within that shift. Appearance, camera, composition, and unseen-operation shifts are evaluated separately.

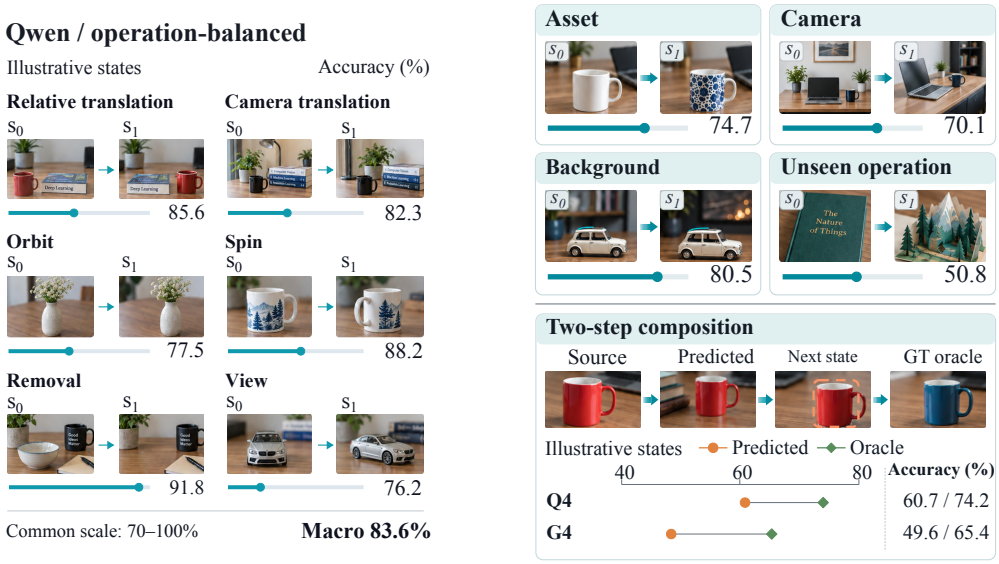
Shift	Method	Translations		Object edits			View	Macro
		Rel.	Cam.	Orbit	Spin	Remove		
ID	SFT	65.7	60.7	54.7	67.7	71.7	55.7	62.7
	Append	70.8	66.8	59.3	73.6	77.3	60.8	68.1
	Replace	**79.6**	**73.6**	**69.0**	**80.7**	**86.4**	**67.9**	**76.2**
Asset	SFT	59.6	52.6	46.6	61.6	66.6	46.6	55.6
	Append	63.9	57.2	49.5	66.4	71.3	49.9	59.7
	Replace	**69.3**	**60.6**	**52.9**	**71.8**	**77.7**	**54.1**	**64.4**
Camera	SFT	57.9	48.9	41.9	60.9	66.9	40.9	52.9
	Append	65.4	55.2	47.4	69.1	75.4	47.5	60.0
	Replace	**70.9**	**59.1**	**51.0**	**73.6**	**82.3**	**51.9**	**64.8**
Background	SFT	63.3	53.3	46.3	67.3	74.3	45.3	58.3
	Append	69.7	58.9	51.4	73.5	81.0	50.1	64.1
	Replace	**78.0**	**66.8**	**58.9**	**81.9**	**90.4**	**57.8**	**72.3**
Composition	SFT	48.6	38.6	30.6	51.6	59.6	32.6	43.6
	Append	54.8	43.7	**34.8**	58.0	65.6	**37.7**	49.1
	Replace	**56.0**	**45.2**	31.0	**59.1**	**69.8**	36.5	**49.6**
Unseen operation	SFT	45.8	38.8	31.8	47.8	53.8	32.8	41.8
	Append	**49.1**	**40.6**	**33.7**	**50.9**	57.7	**35.0**	**44.5**
	Replace	49.0	40.3	29.6	50.6	**59.4**	33.3	43.7

H.2 Intermediate states and recomputation

Table H.3 compares a predicted intermediate state, a true intermediate state, and full-source recomputation. Qwen improves from 60.7% to 74.2% with the oracle intermediate state; recomputation reaches 68.9%. Gemma's predicted-chain result is only 0.5 points above append and also benefits from the oracle. Figure 5(c) illustrates how a wrong intermediate relation propagates into the second edit. The scene sequence is an explanation of the mechanism, not a decoded latent trajectory.

Table H.3: **Composition diagnostics (accuracy, %).** Two-step endpoints share the composition population; true intermediate banks are oracle inputs. Three-step requests remain a separate unsupported population using native fallback.

Model	Chain protocol	Accuracy (%) ↑
Qwen3-VL-4B	Two-step: SFT	52.1
	Two-step: Append	58.0
	Two-step: predicted intermediate	60.7
	Two-step: oracle intermediate	74.2
	Recompute from source + full edit sequence	68.9
	Three-step: native fallback	42.8
Gemma-3-4B	Two-step: SFT	43.6
	Two-step: Append	49.1
	Two-step: predicted intermediate	49.6
	Two-step: oracle intermediate	65.4
	Recompute from source + full edit sequence	59.1
	Three-step: native fallback	36.1



(a) Operation semantics (b) Shift and composition

Figure H.1: **Successful single edits do not guarantee transfer or composition.** (a) Six operation classes have distinct transformation semantics and operation-balanced accuracies. (b) Appearance, camera, and operation shifts are separated from the predicted/oracle intermediate-state comparison. Images illustrate transformations; numerical results summarize the stated populations.

Recomputation is not equivalent to putting unchanged old coordinates into a post-state. It applies the edit again using the source, which lies outside the raw-averaging exclusion in Lemma B.3. Its stronger composition result shows when retaining a route to the original evidence is useful.

H.3 Source-retention policies

Table H.4 uses public-input base-model features and development-defined risk groups. Features include edit class, confidence, parseability, response consistency, invariant-fact stability, and changes in state descriptions. Neither test answers nor trained replacement gains define the groups. Their test shares need not be equal.

Table H.4: **Risk-group contributions.** Observed group weights and replacement-minus-append accuracy (pp) are reported separately for each family.

Family	Risk group	Gain (pp)	Share (%)
Qwen3-VL-4B	Low	6.1	40.0
	Middle	11.5	35.0
	High	21.3	25.0
Gemma-3-4B	Low	4.6	35.0
	Middle	10.0	40.0
	High	18.8	25.0

Replacement-minus-append gains increase from 6.1 to 21.3 points for Qwen and from 4.6 to 18.8 for Gemma. Weighting the unrounded group gains by their reported shares recovers the native differences, 11.8 and 10.3 points. The weighting connects the aggregate gain to observable heterogeneity; the policy experiment below tests whether that grouping is actionable.

Table H.5 tests policy value directly. A quality/conflict gate improves on fixed choices in the balanced construction/conflict population, but gives 57.1% on native MindEdit, below always-replace at 57.8%. The oracle upper bound uses action outcomes unavailable to deployment. Good subgroup separation is therefore insufficient to establish a better real-distribution policy. Figure H.2 separates risk-group differences from policy accuracy. Increasing replacement gains across the groups identify

heterogeneous evidence value; selecting an action must additionally overcome prediction errors under the native mixture.

Table H.5: **Evidence-use policies.** Accuracy (%) ↑ and regret to oracle action selection (pp) on balanced and native mixtures. The public-input gate uses predicted construction quality and edit conflict.

Population	Policy	Accuracy (%) ↑	Regret (pp)
Balanced quality/conflict	Always append	67.8	11.5
	Always replace	68.4	10.9
	Entropy gate	68.8	10.5
	Public quality/conflict gate	73.1	6.2
	Oracle action selection	79.3	0.0
Native MindEdit	Always append	46.0	13.7
	Always replace	57.8	1.9
	Entropy gate	52.6	7.1
	Public quality/conflict gate	57.1	2.6
	Oracle action selection	59.7	0.0

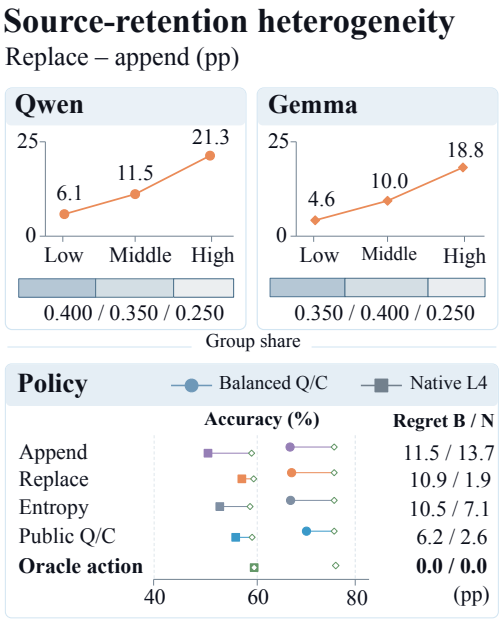


Figure H.2: **Retention policies depend on the prevalence of quality–conflict groups.** The diagram illustrates the groups used to evaluate source availability. The public-input policy improves a balanced mixture but remains below always-replace on native MindEdit. Table H.5 reports policy accuracy separately from the diagnostic group contrasts.

The balanced population deliberately exposes poor and conflicting states. Native requests have different prevalences, so gate errors can outweigh the benefit available in those diagnostic groups. Identifying quality and conflict as relevant variables does not by itself make them reliably estimable from a public input.

I QUERY AND TRANSITION SUFFICIENCY

State compatibility identifies the world to which evidence applies; sufficiency asks whether that evidence determines the requested answer. A state may have the right geometry but omit historical correspondence, or support a terminal answer without supporting another edit. The following comparisons distinguish information that a better representation can preserve from information that the allowed inputs never supplied.

Table 4 groups the retained controls by evaluation setting. Fixed-bank donor comparisons hold construction and reader constant; the other rows address distinct information regimes and use their own populations. Additional geometry labels and experts remain outside the paired-data comparison.

I.1 GEOMETRY, MARGINS, AND SUBSEQUENT EDITS

Construction quality alone cannot specify a useful representation. A rigid spatial query can be answered through explicit geometry, whereas a near-boundary relation depends on coordinate errors in a particular direction. Table I.1 extends the main regime comparison with symmetric orientation and near-margin queries. Symmetry provides no dependable directional cue; a narrow answer margin allows a small state error to change the decision. The reader-stability condition in Appendix B.6 makes the latter dependence explicit through both the representation error and the reference answer margin.

Table I.1: **Additional Qwen3-VL-4B boundaries (accuracy, %).** The other nine regime endpoints are in the main boundary table. Near-margin and symmetric cases test different sources of instability.

Regime	Paired-data interface		Additional information	
	Append	**Replace**	**Text (P+G)**	**Geom. (E)**
Near query margin	57.6	56.8	62.9	66.4
Symmetric orientation binary	49.9	50.0	49.8	50.0

The low-conflict regimes expose a different limit. With no edit or an already compatible input, little obsolete displacement remains to remove. Append and replacement can consequently be close even when both are accurate. In the poor-construction regime, by contrast, the source can restore information missing from the constructed state. These conditions distinguish absence of a useful intervention from failure to construct or read the required information.

A second edit changes the sufficiency requirement again. The current state must preserve the objects, reference frames, and relations that the next transformation reads, including details that the current terminal question may ignore. In Appendix H.2, replacing a predicted intermediate with an oracle intermediate improves composition, while recomputing from the original scene recovers part of the loss. These interventions act at different points: one supplies a better intermediate, and the other retains access to the information needed to reconstruct it. The geometry and composition panels of Fig. 5 therefore test two uses of the same state, answering now and transforming later.

I.2 HISTORICAL CORRESPONDENCE AND AMBIGUITY

A question about the object that was originally leftmost cannot be reduced to its final coordinates without preserving which object moved. Table I.2 tests that missing correspondence. Forced final-state-only reading gives 26.4% on the four-way historical task; declared fallback reaches 48.2%, and a predicted identity map reaches 67.9%. History-preserving text supplies the same kind of information through a different representation. The 97.6% oracle-correspondence result locates the remaining gap after that information is supplied.

Table I.2: **Historical identity and hidden information (accuracy, %).** Historical identity is a four-way task; identical-input hidden pairs are balanced binary tasks. The revealed-information and identity-correspondence oracles change the information supplied.

Diagnostic	Evidence	Accuracy (%) ↑
Historical 4-way identity	Forced final-state only	26.4
	Declared native fallback	48.2
	Final state + predicted identity map	67.9
	History-preserving predicted text	68.4
	Oracle identity correspondence	97.6
Identical-input binary pair	Append; invalid0.6%	49.7
	Replace; invalid0.8%	49.6
	Text; invalid0.4%	49.8
	Geometry; invalid1.0%	49.5
Revealed hidden information	Extra-information oracle	94.3

Figure 5(b) illustrates the distinction between final geometry and a source-to-target identity link. Keeping this link is compatible with an edited-state representation: it preserves a referent across time rather than reinstating an obsolete relation. Source removal is therefore an evidence-interface choice, not a requirement to erase all historical information.

Identical-input binary pairs impose an observational limit instead of a representation-design problem. Ordinary methods remain near 50%; revealing the hidden information raises accuracy to 94.3%. Lemma B.4 explains the ceiling when both possible worlds have identical allowed inputs and different answers. Invalid outputs can lower a measured score below that ceiling. These constructed pairs diagnose missing information, rather than estimate its prevalence in either benchmark. Together, the two comparisons distinguish recoverable history from ambiguity that no transformation of the supplied input can resolve.

J Computation and Cost

A constructed state can be shared across terminal questions, but answering each question still requires reading. Efficiency therefore depends on which work belongs to the scene and edit and which work belongs to the query. The profile separates those stages before evaluating reuse and a larger generation allowance.

J.1 Construction and end-to-end latency

Table J.1 reports the measured mean stage times on one H200 at batch size one. Preparation comprises parsing, image encoding, and edit encoding; state construction applies the transformation; reading comprises prefill and decoding. The six stage means sum to the mean request time. Replacement takes 1.18 seconds, compared with 1.42 for append and 1.02 for paired SFT. It adds construction work to paired SFT but avoids reading the extra source banks presented by append. Pixel generation and adaptive-view methods include the work of their generation or external expert route.

Table J.2 reports end-to-end quantiles and peak allocated memory. These quantities describe complete requests in the same profiling configuration: replacement has a 1.21-second median and a 1.99-second 95th percentile. Quantiles are measured over request times, rather than obtained by adding stage quantiles.

Table J.1: **Mean end-to-end profiling time on a single H200 with batch size one (seconds).** Image/edit columns are encoding, and state/expert includes construction or additional generation. Stage means sum to the total; quantiles are reported separately.

Method	Preparation (s)			State / expert	Reading (s)		Total
	Parse	Image	Edit		Prefill	Decode	
Paired SFT	0.00	0.21	0.00	0.00	0.18	0.63	1.02
Replacement	0.11	0.21	0.04	0.06	0.18	0.58	1.18
Append	0.11	0.21	0.04	0.06	0.32	0.68	1.42
Predicted text	0.11	0.21	0.04	0.37	0.12	0.51	1.36
Geometry executor	0.11	0.68	0.04	0.02	0.12	0.48	1.45
Pixel-state generation	0.11	0.21	0.04	90.00	0.18	0.64	91.18
Adaptive world-model	0.11	0.21	0.04	15.60	0.18	0.66	16.80

Table J.2: **End-to-end latency quantiles and peak allocated GPU memory.** Quantiles are measured on complete requests; memory denotes allocated rather than reserved storage.

Method	End-to-end latency (s)			Memory
	Mean	Median	95th	Peak (GiB)
Paired SFT	1.02	1.07	1.78	18.4
Replacement	1.18	1.21	1.99	21.2
Append	1.42	1.44	2.43	23.8
Predicted text	1.36	1.39	2.62	22.1
Geometry executor	1.45	1.49	2.71	24.7
Pixel-state generation	91.18	89.50	151.00	43.5
Adaptive world-model	16.80	15.90	31.20	52.8

J.2 Reuse across queries

For a fixed scene and edit, cached construction avoids repeating query-independent work (Table J.3). Four queries take 3.46 seconds with reuse and 4.72 without it, including query-dependent prefill and decoding in both totals. Cached and uncached states give the same per-query accuracy.

Table J.3: **State reuse across questions.** Total time in seconds and accuracy (%) ↑ for cached and uncached construction. Joint accuracy requires every answer for the scene to be correct.

Queries	Total time (s)		Per-query accuracy (%) ↑		Joint (%)
	Cached	No cache	Cached	No cache	All queries
1	1.18	1.18	76.0	76.0	76.0
2	1.94	2.36	76.0	76.0	60.6
4	3.46	4.72	76.0	76.0	42.2
8	6.50	9.44	76.0	76.0	22.6

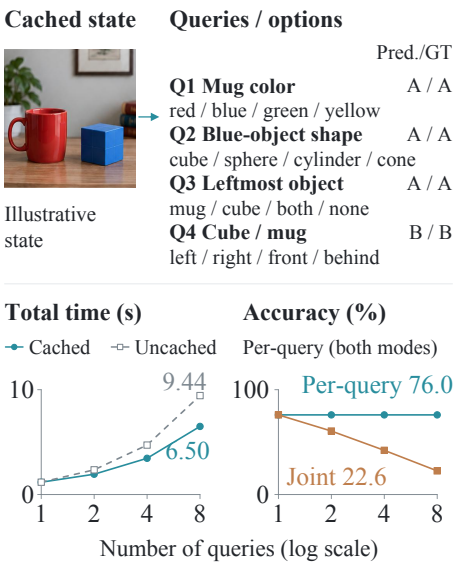


Figure J.1: **Reusing a constructed state saves repeated construction.** Cached and uncached reading have equal per-query accuracy. Joint accuracy requires every answer for a scene to be correct, so it decreases as more queries are included. Timings include separate terminal reading for each query.

Joint accuracy requires every query for a scene to be correct. It falls from 76.0% for one query to 42.2% for four despite unchanged per-query accuracy (Fig. J.1). This stricter event, not a changed cached state, explains the decline; reuse saves construction without making reading decisions identical.

J.3 Answer-generation budget

Table J.4 varies the reader cap over 128, 512, and 2,048 tokens. Raising it from 512 to 2,048 improves Qwen replacement from 57.8% to 58.5% and paired SFT from 33.1% to 36.2%. Append also improves but does not close the replacement gap.

Table J.4: **Reader-token-cap comparison on complete MindEdit-L4 (accuracy, %).** The main configuration allows 512 tokens. Caps specify the generation allowance, rather than realized token count or equal wall-clock work.

Token cap	Qwen3-VL-4B			Gemma-3-4B		
	SFT	**Append**	**Replace**	**SFT**	**Append**	**Replace**
128	31.0	43.5	55.6	26.2	37.0	47.8
512	33.1	46.0	57.8	28.7	39.9	50.2
2048	36.2	47.4	58.5	30.9	41.8	51.1

Each request returns one answer and stops by its cap. A larger allowance adds reading work; reuse removes repeated construction. These inference profiles exclude training compute.

K Limitations

The study concerns object-level visual counterfactual reasoning under explicit evidence interventions. Native verification, endogenous replay, and supplied-card reading establish complementary results but have different histories and populations. In particular, a card-conditioned rate is not an occurrence rate over all native requests. An observable description is also not a complete measurement of an internal world model.

The coordinate analysis is exact within a restricted response class. Reapplying an edit, nonlinear reasoning, and responses outside the candidate–source line can repair cases that raw mixing cannot. Answer probabilities further require a joint margin distribution and calibration, beyond the three risk

moments. Composition and zero-calibration family transfer delimit the predictive instrument even where joint errors remain informative. Finite dose grids identify brackets rather than exact continuous crossing points.

State compatibility makes evidence applicable to the queried world; it neither guarantees sufficient information nor universally requires source removal. Poor construction benefits from a correction route. Historical identity requires correspondence, explicit rigid geometry can favor execution, and an inaccurate intermediate state can break composition. The failed native gating result shows that identifying relevant variables does not automatically provide a better public-input policy. The evaluated pretrained families may also share training data; these comparisons establish neither absence of pretraining contamination nor universal architectural transfer.

Mistaking a hypothetical scene for an observed one can lead to misleading spatial guidance. Observed and queried states should remain explicitly distinguished, and missing information should not be treated as recovered geometry. The experiments do not validate safety-critical physical control. Reproduction and distribution of derived assets must respect the original data and model licenses.