ANSWERS ARE NOT ROUTES: WHAT LOCAL AUDITS ESTABLISH UNDER VISUAL SUPPORT CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can answer correctly and pass local tests while still depending on visual context unrelated to the task. When admissible inputs must satisfy constraints, those tests can omit the comparisons that would expose this dependence. We address this problem with a support-auditing procedure built on the Intervenable Causal Route Bottleneck (I-CRB), an interface for controlling object inputs. For each locally tested task pair, the audit completes its legal contexts and checks whether predictions agree when task inputs are unchanged. It returns counterexamples, completed support checks or unresolved comparisons, with finite-domain guarantees and bounds for partial queries. In a constrained four-object task, we show that even a one-object shortcut can evade every legal one-object context change. Evaluating a task-adapted VLM on 256 new images, our audit uncovers previously undetected context dependence in ten of the 234 images that pass all local tests. Donor-free corrections raise image-equal full-domain accuracy from 97.04% to 97.82–98.34%, with net repair in all twelve runs. Re-auditing identifies remaining support violations; matched comparisons do not establish a mechanism-specific repair advantage. Our approach makes hidden context dependence testable and tracks both accuracy and visual support after model updates.

1 INTRODUCTION

Vision–language models (VLMs) can answer a visual question correctly while depending on objects that do not determine its answer. Testing nearby inputs appears to offer stronger evidence: change one object at a time and check that the model still answers correctly. Yet even passing every allowed local test can leave this unwanted dependence undetected. The revealing comparison may lie outside the allowed local changes. This gap matters when using tests to judge whether a model relies on the intended visual evidence, or whether a repair has corrected that behavior.

Existing approaches examine visual evidence at different levels. Concept and clue bottlenecks make predictor inputs explicit (Koh et al., 2020; Fu et al., 2024). Counterfactual visual question answering changes inputs to test which evidence affects an answer (Agarwal et al., 2020; Chen et al., 2020), while visual feature supervision separates accuracy from sensitivity to relevant and irrelevant features (Ying et al., 2022). Internal interventions replace representations to test their effects on answers (Geiger et al., 2021; 2024). Constrained testing already shows that admissibility limits which explanations can be distinguished (Jin & Tsuchiya, 2020).

The challenge is to determine what a successful visual audit establishes when its comparisons are constrained. In our four-object task, spatial relations restrict object selections, although only the first two objects determine the answer. Replacing the fourth may be illegal until the third changes. An endpoint-only rule can therefore pass every allowed one-object context edit and fail after two edits (Figure 1). The missed dependence is simple; reaching a revealing comparison is not. This motivates our question: which claims about visual support can available tests establish, and do those claims survive a model update?

We address this question with a support-auditing procedure built on the Intervenable Causal Route Bottleneck (I-CRB), which controls the object inputs supplied to a predictor. A "route" denotes their ordered selection, not an internal neural pathway. After local testing, we complete all legal contexts

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.

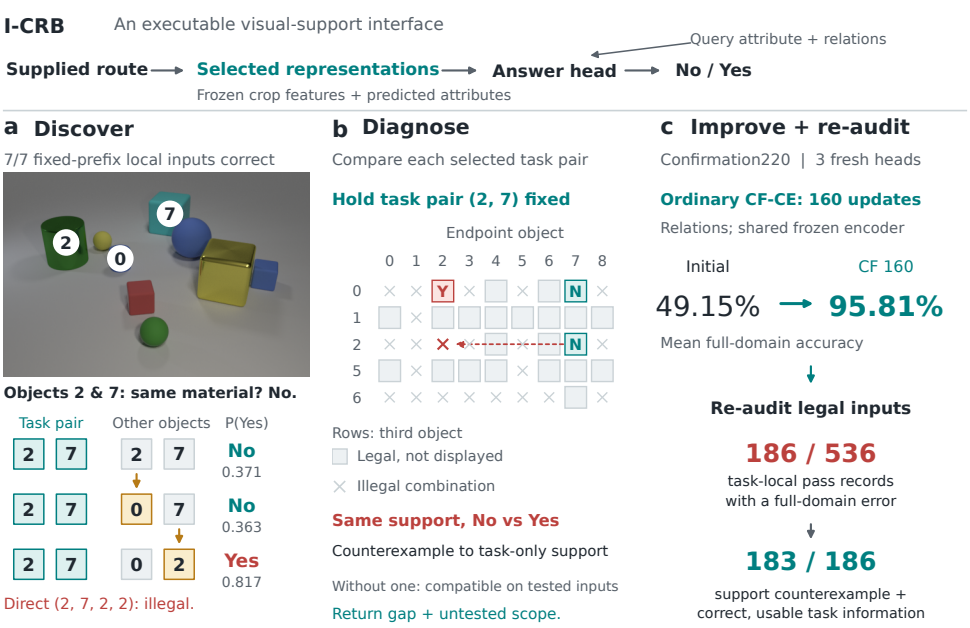


Figure 1: **From a missed comparison to diagnosis and re-audit. a**, In scene 046684, replacing the last object is illegal until the third selection changes. An endpoint-dependent answer can therefore survive the immediate checks. **b**, The audit completes legal contexts at each tested task pair. The feature/head diagram depicts the controlled model, not the complete VLM. **c**, On a separate improved-head panel, 92/536 local passes fail completion; 186/536 fail the full domain. Of these, 183 have a same-prefix contradiction at a wrong input with correct task attributes and Lookup/Restricted answers. Counts concern model–image records containing errors, not wrong inputs. Arrows describe the audit, not a neural circuit.

for each tested task pair. Different predictions for unchanged task inputs reject task-only support. Otherwise, the audit reports completed checks and comparisons still missing.

Our contributions follow the progression from discovery to diagnosis and improvement followed by re-auditing. We establish that legal reachability and dependence complexity differ: even a one-feature shortcut can require multiple legal edits to expose. We provide an executable support audit with counterexamples, finite-domain certificates and sharp bounds when comparisons remain unqueried. Finally, controlled models and a task-capable VLM demonstrate why accuracy gains must be evaluated alongside support behavior. Frozen corrections yield net repair on new images, while re-auditing identifies residual dependence that local tests still miss. Matched controls separately assess whether intervention-trained directions add repair benefit, distinguishing semantic control from repair effectiveness.

2 What We Test and What a Pass Means

2.1 A Four-object Task with Known Relevant Inputs

We use CLEVR scenes (Johnson et al., 2017) because object attributes and spatial relations are available for checking each intervention. A route $r = (v_0, v_1, u, w)$ selects four objects satisfying three directed spatial relations. The question asks whether the first two have the same material or size:

$$T(r) = \mathbf{1}\{a(v_0) = a(v_1)\}. \tag{1}$$

Here a is the queried attribute and T is the correct binary answer. We call (v_0, v_1) the *task pair* (prefix) and (u, w) the *suffix*. Changing the suffix leaves the target unchanged. For scene object set V and a fixed relation program (R_0, R_1, R_2), the legal domain is $\mathcal{D} = \{r \in V^4 : \bigwedge_{j=0}^{2} R_j(r_j, r_{j+1})\}$. Relations come from CLEVR annotations, not estimates from the montage. The Relations domain

Table 1: **How to read an audit result.** Support checks compare predictions; correctness checks also require task labels. The input set and unqueried remainder are always reported.

Result	What was observed	What follows
Counterexample	Same proposed support, different answers	That support alone cannot explain the predictions
Completed check	All legal inputs in a chosen group were queried; no contradiction	Compatible within that group, not necessarily correct
Unresolved check	Some legal comparisons remain unqueried	No conclusion about those missing comparisons
Re-audit	Same inputs evaluated before and after correction	Old errors repaired, errors retained and new errors introduced

permits repeated objects when the three tests hold; All-distinct additionally requires four different objects. Relations constrain selections, not equality itself; no objects move. This experimental contract isolates restricted comparisons, without establishing their necessity or prevalence in deployment.

I-CRB exposes these selections at the model input. The controlled model feeds frozen crop features and predicted attributes to a two-layer Transformer head. Lookup compares predicted attributes; Restricted learns from task positions alone. Ground truth supplies targets only. For a complete VLM, the Four interface arranges the crops as labeled tiles and asks whether the centered objects in tiles 1 and 2 match. Pair supplies only those two crops. Support concerns selected input positions, not perfectly isolated semantics: crops may contain nearby objects. Appendix B gives the interfaces, training and capability checks.

2.2 Local Tests and the Comparisons They Leave Out

Fix the scene, question and spatial relations, and start from an original route r_0, also called the root. The local audit $\mathcal{L}_{\text{task}}$ contains this root and every legal change to one selected object. It includes changes to task objects: their correct answers are recomputed. A *local pass* means that every answer in this set is correct.

A wider audit separates context changes for tested task pairs from generalization to task pairs never reached locally. Let $\mathcal{D}$ be all legal routes, $\pi(r) = (v_0, v_1)$ extract the task pair, and $\mathcal{P}$ collect the pairs reached locally. Their *selected-prefix completion* is

$$\mathcal{P} = \pi(\mathcal{L}_{\text{task}}), \qquad \mathcal{U}^\star = \{r \in \mathcal{D} : \pi(r) \in \mathcal{P}\}. \tag{2}$$

Thus $\mathcal{U}^\star$ completes contexts at locally tested task pairs, while $\mathcal{D}$ also includes untested pairs. Earlier experiments use the original-pair set $\mathcal{F}$ and $\mathcal{U} = \mathcal{F} \cup \mathcal{L}_{\text{task}}$, giving $\mathcal{F} \subseteq \mathcal{U} \subseteq \mathcal{U}^\star \subseteq \mathcal{D}$. The distinction concerns which comparisons are available, not just the number of inputs.

2.3 Returning Evidence, Not Just a Pass/Fail Score

A support hypothesis says that the prediction depends only on specified input positions. For example, task-only support uses positions $J = \{0, 1\}$; endpoint-only support uses $J = \{3\}$. To test either hypothesis, group queried routes by their selections at those positions. Two different answers within a group are a counterexample: those positions alone cannot explain both predictions.

Formally, let $\mathcal{H}_J$ contain all deterministic functions of the ordered selections r_J, with the scene, question and relations fixed. Predictions $\hat{f}$ on queried inputs $\mathcal{Q}$ are compatible with this family exactly when

$$r_J = s_J \implies \hat{f}(r) = \hat{f}(s) \quad \text{for all } r, s \in \mathcal{Q}. \tag{3}$$

A function can then assign the common answer to each group, but need not be correct or remain compatible on unqueried inputs. We apply this classical criterion (Skowron & Rauszer, 1992) to enumerated legal visual comparisons (Table 1).

The decisive information is a correct answer at the *same* task pair, which we call an anchor.

Table 2: **A support certificate changes the conclusion, not just the score.** Scene 001514, size, unrepaired Four. Seven local queries are correct, but 29 of 36 selected-prefix completion inputs remain unqueried. The correct anchor $(2, 4, 0, 2)$ answers No; the legal completion $(2, 4, 0, 4)$ answers Yes for the same task pair and target No. This is the first completion miss in image-name order. Gap bounds concern all binary extensions, not confidence intervals.

Observed queries	Missing	Gap range (%)	Support judgment
Seven local inputs	29	[0,50]	Unresolved
Local + one contradictory completion	28	[2.78,50]	Task-only rejected
All 36 completion inputs	0	[8.33,8.33]	Task-only rejected

Proposition 1 (What completion certifies)**.** *Suppose the target depends only on the task pair and every local input is answered correctly. Task-only support on $\mathcal{U}^\star$ then holds if and only if every input in $\mathcal{U}^\star$ is answered correctly. Without additional assumptions on the predictor, certifying either property requires all $|\mathcal{U}^\star \setminus \mathcal{L}_{\mathrm{task}}|$ missing queries in the worst case.*

Every selected pair has a correct local anchor. A wrong completion contradicts that anchor; an unqueried completion could conceal just such a contradiction. Appendix C.3 proves the statement and gives the corresponding random-sampling bound. An error on an untested pair in $\mathcal{D} \setminus \mathcal{U}^\star$ lacks the correct anchor and does not, by itself, reject task-only support. We also report the fraction of predictions that must be discarded to remove contradictions, called the support gap (Parciak et al., 2025). Appendix C defines it and gives sharp bounds when some comparisons remain unqueried. These are logical identification bounds, not statistical confidence intervals or predicted error risks. Table 2 shows an actual decision change. Of 234 VLM local passes in the new-image confirmation, 232 remain unresolved before completion. Afterwards, ten contradict task-only support and 224 are correct throughout $\mathcal{U}^\star$. The certificate prevents unsupported certification; it does not rank those 232 images by risk or maximize errors found per query (Appendix E).

3 Discovery: Why Local Tests Miss Simple Errors

3.1 One-feature Dependence Can Require Two Legal Edits

Consider answering Yes whenever the last object is predicted metal. If every immediately reachable endpoint has the same cue, one-edit suffix tests cannot reveal the shortcut. Changing the third object may then permit a revealing endpoint without changing the task pair (Figure 1).

The exact condition is simple to state. Write the endpoint rule as $g(r) = \widehat{c}(w)$, where the cue is predicted metal for material or predicted large for size. Let $\mathcal{A}$ contain third objects allowed by the fixed prefix, and $\mathcal{B}(u)$ contain their legal endpoints, and let $\mathcal{L}_{\mathrm{suffix}}$ contain the root and all legal one-edit suffix changes. For a correctly answered root, with target $y = T(r_0)$,

$$g \text{ passes } \mathcal{L}_{\mathrm{suffix}} \text{ but fails } \mathcal{F} \iff \begin{aligned} &\forall w \in \mathcal{B}(u_0),\ \widehat{c}(w) = y, \\ &\exists w \in \bigcup_{u \in \mathcal{A}} \mathcal{B}(u),\ \widehat{c}(w) \neq y. \end{aligned} \tag{4}$$

All immediate endpoints look safe, but another legal endpoint is wrong. Changing only u leaves this rule's answer unchanged. In the original spatial setting, any two legal suffixes can be joined in at most two one-object edits (Appendix C). Two edits therefore need not mean dependence on two features.

On 128 discovery images, the endpoint rule shares 65 of the learned head's 66 fixed-prefix misses. Task-object edits reveal 55 of those 66, so the stronger local baseline removes many blind spots, but not all. After training, some local passes correctly handle target-changing edits that an endpoint-only rule cannot handle. The rule explains the initial regime, not every later error.

3.2 The Blind Spot Persists after Accuracy Improves

Counterfactual cross-entropy (CF-CE), ordinary supervision on original inputs and legal alternatives, improves the controlled model. At 160 matched updates, $\mathcal{U}$ accuracy reaches 98.19%, versus

Table 3: **Improvement does not remove the need for a wider audit.** 220 new images and three fresh answer heads after 160 CF updates. Each fraction counts model–image records containing an error among locally passing records, not wrong inputs. Relations uses the spatial rules; All-distinct also requires four different objects.

Confirmation legality	$\mathcal{D}$ acc.	$\mathcal{F}$ miss/pass	$\mathcal{U}^\star$ miss/pass	$\mathcal{D}$ miss/pass
Relations	95.81	10/536	92/536	186/536
All-distinct	94.79	6/563	59/563	185/563

Table 4: **A task-capable VLM still leaves untested errors.** Unrepaired task-adapted Qwen, image-equal accuracy. Each row is a different panel; denominators are locally passing images. The first row used batched inference, with 4/5 completion misses surviving batch-one replay. Other rows use batch one. The last two panels are development data, not new confirmations.

Panel	Images	$\mathcal{D}$ acc. (%)	$\mathcal{U}^\star$ miss/pass	$\mathcal{D}$ miss/pass
Initial audit	128	97.34	5/119	12/119
Old development	128	96.92	2/117	10/117
New development	64	96.75	3/59	7/59

54.93% for repeated originals and 57.52% for cue-balanced originals. Matched initialization, order, optimizer reset and batch size support the recipe comparison, not coverage as its sole cause.

On 220 further images and fresh heads, 17.16% [13.87,20.63] of Relations local passes fail completion (Table 3; fixed-head, image-cluster 95% interval). Full-domain failures additionally include untested task pairs.

All 19 wrong inputs in the older anchor domain have correct task attributes and Lookup/Restricted answers. Isolated-object rendering also leaves completion misses in a 99.48%-accurate head. These controls limit specific perceptual explanations; Appendix D retains both and independent two-task CF repair.

3.3 A Task-capable Complete VLM Also Passes Locally and Fails Later

Native-frozen models fail the task-accuracy checks, so their exploratory audits do not establish the phenomenon for capable models (Appendix B). We instead adapt the complete Qwen3-VL-30B-A3B-Instruct model (Qwen Team, 2025) using low-rank adaptation (LoRA) on 512 original Four inputs. It answers 127/128 Four and 125/128 Pair development inputs correctly. This is a task-adapted mixture-of-experts VLM, not native zero-shot validation. Table 4 reports aggregate audits. Four of the initial five completion misses survive batch-one replay; one is batch-size-sensitive. Subsequent mechanism and correction experiments use batch size one.

A constraint-strength check holds Four, the root, question and original completion reference fixed, while removing only the final spatial relation. The 725 additional one-edit inputs expose one of five previously hidden completion errors across 192 development images. An equal-query analysis separates access to that comparison from simply making more calls (Table 5). Relaxation makes the focal contradiction discoverable at the same cost, but does not improve overall discovery under uniform sampling. Thus constraints can exclude useful comparisons without explaining every blind spot or being necessary for the task itself.

4 Diagnosis: What Changes inside the Model?

An intervention can control an answer without explaining its natural error (Makelov et al., 2023; Hase et al., 2023). Figure 3 connects three tests: state replacement, semantic control and natural-error transfer.

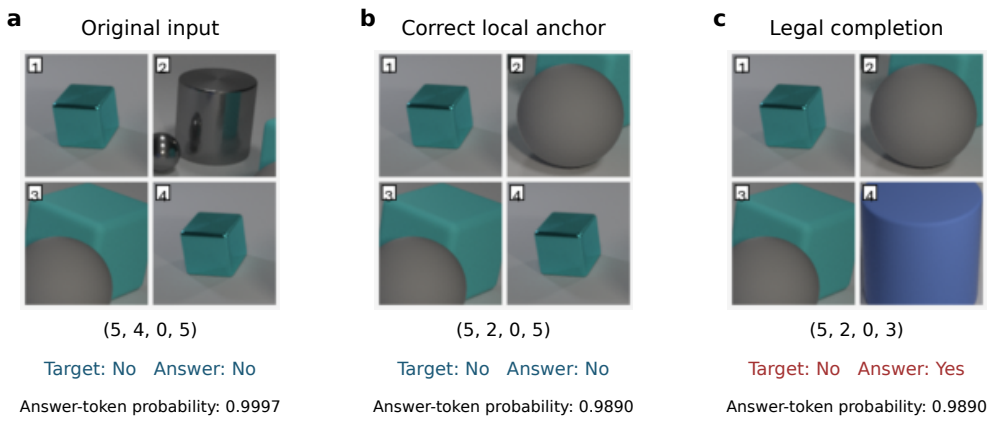


Figure 2: **The same task images, a different answer.** Scene 011542 asks whether the centered objects in tiles 1 and 2 have equal size. All eight local queries are answered correctly. A legal completion for task pair $(5, 2)$ keeps those two tiles and the target unchanged, but changes the model's answer. This counterexample is the first residual in the fixed panel order and survives batch-one replay. Probabilities are normalized over No/Yes. The example demonstrates dependence beyond the task tiles; it does not identify a neural layer responsible for the error.

Table 5: **More admissible comparisons need not improve uniform-query efficiency.** Post hoc exact analysis on 192 development images, with the original completion reference fixed. The root is shared; each equal-cost policy queries as many nonroot inputs as the strict local set contains. Relaxed sampling is uniform without replacement. Entries are expected numbers of images, not fractional observed cases. "Focal" refers to the five original local-pass/completion-error cases. Each logical query costs one scoring and one generation call.

Query policy	Queries	Errors found	Support conflicts	Focal exposed
Strict local, exhaustive	986	16	8	0
Relaxed local, same budget	986	12.15	6.67	0.68
Relaxed local, exhaustive	1,711	17	9	1

4.1 Testing the States Associated with the Task Objects

Unchanged task pixels can acquire different internal states when encoded with different contexts. We replace task states across the main visual output and DeepStack streams from visual layers 8, 16 and 24. Each stream has 169 positions, with 25 strict-interior positions per tile. For correct run G and incorrect run B with the same task pair, we copy task states from G into B, leaving other positions unchanged. After four development cases, the frozen operation restores eight of ten new-image residuals. Its directional-margin advantage over equal-sized suffix replacement is 1.03 [0.06,5.50] logits. Correct/correct controls damage one answer in 20 directions. The four streams change together: this identifies an effective intervention, not a uniquely responsible layer (Appendix G).

4.2 Reading an Attribute Is Different from Changing Its Use

A linear classifier reads material and size with 98.18% and 99.74% held-out accuracy, yet replacing its directions changes 126/128 decoded attributes and none of the expected VLM answers. We instead learn exchanges targeting the answer implied by the exchanged attributes, following distributed alignment (Geiger et al., 2024). For example, exchanging metal for rubber should change the equality answer according to the other object's material, not copy the donor's answer. Four orthonormal directions $Q_c \in \mathbb{R}^{2048\times 4}$ per stream restrict replacement to a learned subspace:

$$h'_c = h_c + Q_c Q_c^\top (h_c^{\text{donor}} - h_c). \tag{5}$$

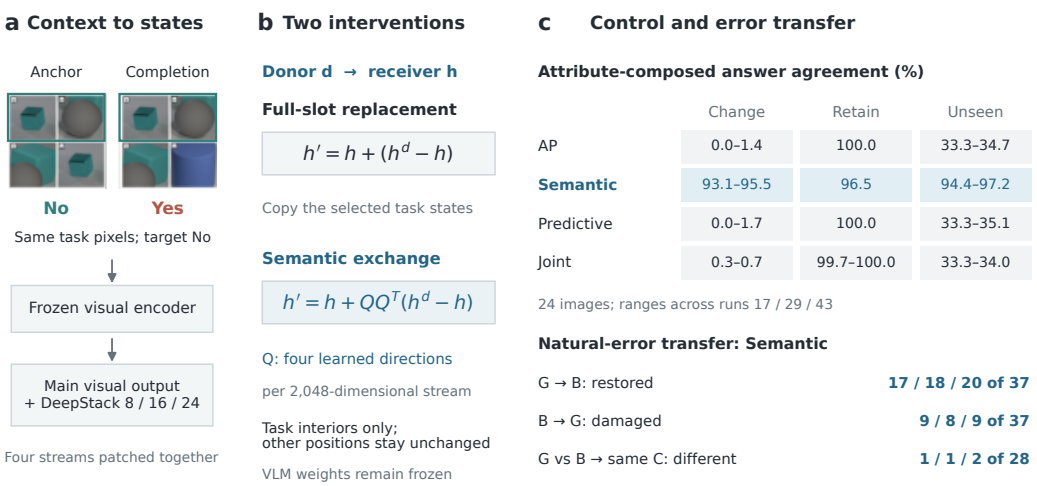


Figure 3: **From a behavioral counterexample to internal intervention tests. a**, The legal anchor and completion from Figure 2 have identical task pixels (outlined) but different answers. The four streams mark intervention access points, not a uniquely causal layer. **b**, Full-slot copies selected states; Semantic exchanges only their rank-four projection. VLM weights and unselected positions stay fixed. **c**, Actual final repair bases, with no refitting. Agreement ranges are minima–maxima across pipelines 17/29/43, not confidence intervals; all share one adapted VLM and 24 test images. Natural-error counts follow the same pipeline order. G/B denote correct/error contexts of the same task pair; C is a fixed third receiver. Restoration, damage and fixed-receiver differences have distinct denominators. Complete controls are in Appendix K.

Receiver and donor tokens occupy matching within-tile positions. Maps are shared across positions but fitted separately for each attribute; other states and VLM weights remain fixed. Construction uses attribute-pair types 00/11; the 24-image test also includes unseen one-attribute changes (01/10). The *answer-preserve* (AP) control shares capacity, initialization and data but targets the receiver's answer. Figure 3c tests the actual final repair bases, not the earlier mapping in Figure 5. Every Semantic basis supports composed-answer control, unlike AP, Predictive and Joint (defined in Section 5). The existing test images are disjoint from construction and repair fitting; no basis is refitted or selected. Appendix K retains all pipeline and attribute results.

4.3 A Successful Exchange Is Not Yet a Full Error Explanation

Restoring a wrong receiver and damaging a correct one do not hold the receiver fixed. Figure 3c therefore also compares correct/error donors in the same 28 third receivers. The final Semantic bases change hard answers in only 1, 1 and 2 cases, although margins differ in 24, 22 and 25. This supports partial error transfer, not a complete natural-error pathway. Historical Full-slot replacement distinguishes seven cases and the earlier Semantic mapping none (Appendix G). Appendix C.6 explains why semantic control alone cannot guarantee natural-error attribution.

5 Improvement: Repair Errors and Check What Remains

5.1 Measure Repair on Fixed Inputs, Not Just Remaining Blind Spots

In one continuation, blind spots fall from five to one, but only two old residuals are repaired: three become local failures and one new blind spot appears (Appendix F). We therefore track fixed-input repairs, persistent errors and *regressions*, originally correct answers made wrong.

5.2 From Internal Exchanges to a Deployable Correction

A frozen donor-selection rule repairs ten errors and introduces nine (Appendix H). We instead predict the update from the current input, keeping the VLM and adapter frozen:

$$h'_{c,t} = h_{c,t} + Q_c\, a_{\phi,c,t}(H). \tag{6}$$

Here $H \in \mathbb{R}^{4\times169\times2048}$ contains the current visual states and $a_{\phi,c,t}(H) \in \mathbb{R}^4$ gives update coefficients. We write at the 25 strict-interior positions of each task tile, 50 positions per stream; the other 119 stay unchanged. This form builds on representation tuning (Wu et al., 2024; Bi et al., 2024). We test the value of intervention-trained directions Q_c, not the novelty of low-rank updates. Earlier comparisons found no established benefit from explicit task grouping (Appendix J). We now fix the Generic reader a_ϕ, which reads each token and three learned attention summaries from the unmodified states.

5.3 A Matched Test of Where to Write the Correction

AP learns directions whose exchanges preserve the receiver's answer; Semantic targets the attribute-composed answer (Equation 5). Predictive uses original answers, attributes and composed targets without patched-VLM calls. Joint learns the same reader and rank-four directions together. It has 168,528 parameters per attribute versus 135,760 for fixed-space correction, so it is not capacity-matched.

All arms share reader initialization, examples, order and 512 repair updates at 10^{-3}. Fitted spaces share a construction pool. Repair uses 965 natural inputs from 64 images and relabeled Pixel composites, which are not legal audit inputs. Three paired construction–repair pipelines share one adapted VLM. Repair budgets are matched; total construction costs are not. The shared objective averages the two natural-input binary CE losses, adds 0.3 times the Pixel CE and 0.3 times the larger natural loss, and penalizes update magnitude with weight 0.01. No donor, label or scene annotation is used at deployment (Figure 6; Appendix L).

The development panels contain 128 and 64 images. At full amplitude, corrections produce net repair, but the original AP advancement rule fails its regression and capability limits. Semantic also loses an old-panel material root in one pipeline. Joint passes the baseline-relative checks; its comparison with Semantic has mixed repeat directions and intervals crossing zero. These outcomes motivate freezing a risk-calibrated choice before testing new images, not declaring a winner from development data. Appendices L and N retain the original primary comparison, all endpoints and failure decisions.

5.4 Freeze Risk Calibration before Testing New Images

Does intervention-based construction help under common risk calibration? Without further training, we choose each method's amplitude $\alpha \in \{0, \frac{1}{2}, 1\}$ on the old 128 images, shared across attributes and pipelines. Among candidates passing capability and regression limits, we minimize completion bad-prefix rate. A bad prefix contains a wrong completion; prefixes are weighted equally within each image. Regression is the image mean of the fraction of originally correct inputs made wrong. The point-estimate limits are mean regression at most 0.5%, at most 1% per pipeline, overall accuracy loss at most 0.2 points, and per-attribute capability losses at most one point. This selects $\alpha = \frac{1}{2}$ for Semantic and one for AP, Predictive and Joint (Appendix O).

We then freeze the choice and evaluate 256 new, geometry-selected images outside fitting and historical evaluation memberships. All 6,299 legal inputs remain, including 3,607 completion inputs. The unrepaired VLM makes 191 errors on 47 images: ten material and 37 size, versus only two and ten on the earlier new-development panel. Its 234 local passes contain ten completion misses and 25 full-domain misses. This strengthens image-level evidence, not model replication: the three pipelines still share one adapted VLM.

All four corrections pass these frozen, sample point-estimate checks, and every run has net repair (Table 6). Every regression interval extends above 0.5%, so these checks do not certify population risk below that threshold. Figure 4 shows the primary Semantic–Joint comparison. The comparison fails the prespecified conjunction of bad-prefix superiority, accuracy noninferiority and regression

Table 6: **Frozen correction on 256 independently selected images.** Each method's amplitude is chosen on the old 128 images and fixed across attributes and runs. All 6,299 legal inputs are retained. Rates (%) are image-equal means over three pipelines sharing one adapted VLM. Regression brackets are 95% image-bootstrap intervals. The mean-regression threshold is 0.5% on this panel, not a population safety guarantee. Repair/New are mean input counts; $\mathcal{U}^\star$ misses pool image–pipeline records, not independent images.

Method (α)	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regression [95% CI]	Repair	New	$\mathcal{U}^\star$ miss/pass
Four (0)	97.040	5.358	0	0	0	10/234
AP (1)	98.109	3.123	0.440 [0.189,0.755]	95.67	23.33	11/720
Semantic (0.5)	97.879	3.533	0.272 [0.065,0.546]	77.33	14.67	15/720
Predictive (1)	97.823	3.945	0.311 [0.130,0.555]	72.67	17.67	19/714
Joint (1)	98.341	2.407	0.361 [0.169,0.591]	109.67	24.67	7/726

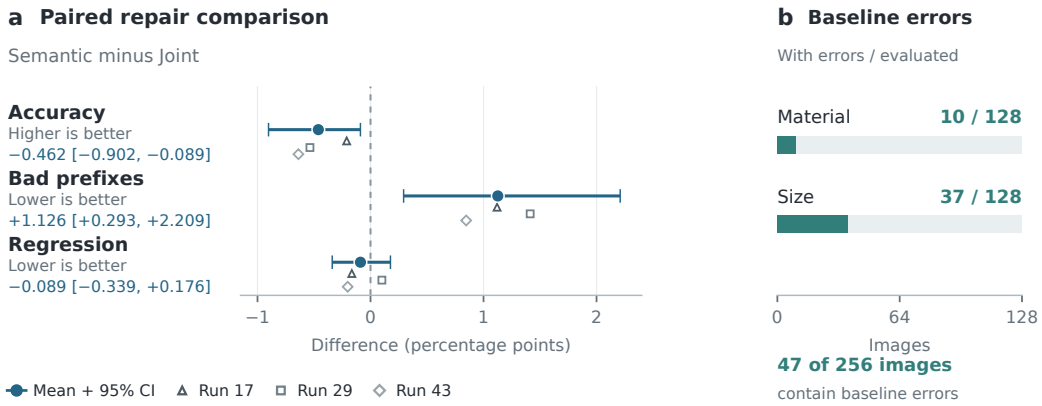


Figure 4: **Net repair does not establish a mechanism-specific advantage.** **a**, Semantic minus Joint on 256 new images. Circles and bars show means and 95% paired-image bootstrap intervals (10,000 draws), averaging pipelines within image. Open symbols show runs 17/29/43, not independent models. Positive accuracy and negative bad-prefix/regression differences favor Semantic. **b**, Bars partition each 128-image attribute panel into images with baseline errors (colored) and without (gray): ten material and 37 size. Model and fits are held fixed.

noninferiority. Joint repairs more errors in all three runs; Semantic's lower mean regression does not establish a safer trade-off. Material also differs from size: Semantic repairs 5.33 material errors but introduces 8.00 on average. The mechanism-specific increment is not confirmed. Calibration does not eliminate capacity or construction-cost differences; Appendix O retains all repeats and transitions.

6 Discussion and Conclusion

Legal reachability and dependence complexity are different: even a one-object shortcut can evade every legal one-edit test. I-CRB completes legal contexts and returns counterexamples, completed support checks or unresolved comparisons, with finite-domain guarantees. Semantic exchanges test answer control; natural-error transfer tests its role in observed errors. Donor-free corrections achieve net repair in every confirmation run under frozen sample risk checks, while mechanism-specific superiority remains unestablished. Re-auditing distinguishes improved answers from resolved support violations.

The rendered scenes, supplied selections and single adapted VLM define our empirical scope. Pair and blanking controls reinforce the need to assess invariance alongside accuracy (Appendix M). Additional models, mandatory constraints and larger independent samples would test transfer and population risk. Together, discovery, diagnosis and re-auditing make visual support an explicit part of model evaluation, before and after repair.

REPRODUCIBILITY STATEMENT

Appendix B specifies the visual interfaces and adaptation; Appendix C gives the support-consistency arguments. Appendices I–L describe correction objectives, selection, stopping and evaluation units; Appendices M–N add same-panel input controls and jointly learned directions. The accompanying source records identify the completed experiments underlying each display. Appendix O specifies the frozen amplitude calibration, 256-image confirmation, relation-removal control and complete resource use. The accompanying anonymous CPU reconstruction archive contains per-input predictions, targets and audit-set membership. The archive reconstructs calibration, geometry, labels, paired-image intervals, support certificates and every primary and safety decision without loading a VLM. Appendix K additionally validates the exact final repair bases; its per-condition outputs and checkpoint identities are supplied. The appendices document the earlier repair comparisons. These are statistical reconstructions, not packages for rerunning training. Cached analysis, new inference and repeated training are distinguished; the new-image confirmation holds one adapted VLM fixed and does not turn earlier development results into confirmation.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting, language polishing and organization, as well as conceptual and mathematical formulation, hypothesis refinement, experimental design, implementation, and result interpretation. The authors are responsible for verifying all AI-assisted content and the final claims.

REFERENCES

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9690–9698, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Agarwal_Towards_Causal_VQA_Revealing_and_Reducing_Spurious_Correlations_by_Invariant_CVPR_2020_paper.html.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA Steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*, 2024. URL https://arxiv.org/abs/2412.12359. Version 2, January 2025.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Counterfactual_Samples_Synthesizing_for_Robust_Visual_Question_Answering_CVPR_2020_paper.html.

Martin C. Cooper and Leila Amgoud. Abductive explanations of classifiers under constraints: Complexity and properties. In *ECAI 2023*, pp. 469–476, 2023. URL https://arxiv.org/abs/2409.12154. Full version with proofs: arXiv:2409.12154.

Xingyu Fu, Ben Zhou, Sihao Chen, Mark Yatskar, and Dan Roth. Dynamic clue bottlenecks: Towards interpretable-by-design visual question answering. *arXiv preprint arXiv:2305.14882*, 2024. URL https://arxiv.org/abs/2305.14882.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings*

of Machine Learning Research, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2301.04213.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. *arXiv preprint arXiv:2402.17700*, 2024. URL https://arxiv.org/abs/2402.17700.

Hao Jin and Tatsuhiro Tsuchiya. Constrained locating arrays for combinatorial interaction testing. *Journal of Systems and Software*, 170:110771, 2020. doi: 10.1016/j.jss.2020.110771.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6781–6792, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023. URL https://arxiv.org/abs/2311.17030.

Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering reliable approximate functional dependencies. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 355–363, 2017. doi: 10.1145/3097983.3098062.

Marcel Parciak, Sebastiaan Weytjens, Niel Hens, Frank Neven, Liesbet M. Peeters, and Stijn Vansummeren. Measuring approximate functional dependencies: a comparative study. *The VLDB Journal*, 34(4):56, 2025. doi: 10.1007/s00778-025-00931-x.

Qwen Team. Qwen3-VL-30B-A3B-Instruct. Official model card and checkpoint, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct. Accessed September 19, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Official Qwen3.5-9B model release; checkpoint revision c202236235762e1c871ad0ccb60c8ee5ba337b9a.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451, 2024. doi: 10.18653/v1/2024.findings-emnlp.312. URL https://aclanthology.org/2024.findings-emnlp.312/.

Andrzej Skowron and Cecylia Rauszer. The discernibility matrices and functions in information systems. In *Intelligent Decision Support*, pp. 331–362. Springer Netherlands, 1992. doi: 10.1007/978-94-015-7975-9_21.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024. URL https://arxiv.org/abs/2404.03592.

Zhuofan Ying, Peter Hase, and Mohit Bansal. VisFIS: Visual feature importance supervision with right-for-the-right-reason objectives. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.11212.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023. URL https://arxiv.org/abs/2309.16042.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A READING THE EVIDENCE AND REPRODUCING THE DISPLAYS

The appendix follows the same questions as the main text. Two historical repair branches are summarized here; the accompanying complete-record PDF preserves their original tables, transitions and implementation detail. No historical arm or stopping decision is removed from the evidence package. Appendix B explains the models and task interfaces. Appendices C–E give the audit arguments, controlled-model evidence and query-budget comparisons. Appendices F–L document the complete-VLM interventions and repair experiments. Table 7 identifies which images support each result.

Two terms are important throughout. "Fresh" means an image was unused in the recorded project evaluations when that experiment began. Once its outputs inform later development, reusing it is not a new independent confirmation. A "trajectory" is a particular training run; three correction trajectories do not mean three independently trained base VLMs. The companion evidence index links experiments to their saved records.

Table 7: Evidence roles and units. All complete-VLM mechanism studies below use the same task-adapted Qwen checkpoint. "Fresh" refers to documented project membership at first inference, not unknown pretraining exposure. Error pairs are subsets, not extra scenes.

Panel / experiment	Images	Role and independence boundary
Controlled discovery	128	New at initial cue-rule discovery; three heads
Controlled extension	512	Opened cue-shift stress panel; three fixed heads
Controlled confirmation	220	Fresh images and three fresh heads sharing one encoder
Independent two-task repair	640	320 new images/task; three fixed visual seeds
VLM matched continuation	128	Fresh when first evaluated; three continuation seeds, one adapter
Visual-interface confirmation	256	Fresh when tested; ten selected-prefix residuals for patching
Slot-composition pilot	128	Fresh; 32 eligible semantic pairs, no local blind cohort
Error-context study	192	Fresh at first evaluation; 19 error pairs, later opened
Learned interchange	48+24	Disjoint map-training/evaluation images from opened material
Final-basis validation	24	Same frozen exchange test, final AP/Semantic/Predictive/Joint bases in three pipelines; no refit
Legal-reference mean	128	Fresh at first inference; 2,923 complete legal inputs
Donor-free repair	64+128	Existing training scenes and the same opened 128-image evaluation panel
Matched write spaces	128+64	Old and new development panels, separate from construction and repair fitting; not final confirmation
Frozen risk calibration	128	Previously opened images only; select one amplitude per method
New-image confirmation	256	Output-blind geometry selection, disjoint from fitting and historical evaluation; one fixed adapted VLM
Constraint relaxation	128+64	725 added one-edit inputs on the same development images; no new scenes

Paper tables and figures report frozen summaries. The source package includes a record-to-display index for the added comparisons. Experiment packages separately preserve protocols, checkpoints, memberships, raw answers and verification scripts. The reference-mean and initial donor-free packages pass clean-directory CPU reconstruction; this is not a second GPU run, independent model fit or external replication. Full model replay requires the named checkpoints, task adapter, CLEVR assets and recorded environment. Internal filesystem paths remain in research archives. The accompanying anonymous CPU archive reconstructs the matched write-space tables and paired intervals directly from stored per-input predictions, targets, margins and set membership. It also reproduces the numerical advancement checks. It does not revalidate GPU training or replace separate model-replay assets. The new confirmation archive additionally reconstructs legality and labels from supplied scene metadata, frozen amplitude selection, support certificates and all confirmation decisions (Appendix O).

B MODELS, INTERFACES AND CAPABILITY

The two model settings serve different purposes. The small CNN and answer head allow direct checks of input information. The complete VLM tests the behavior and internal interventions in a pretrained system. Task adaptation, capability checks and later repair are separate stages, described below.

B.1 CONTROLLED VISUAL MODEL

The frozen encoder receives zero-padded 96×96 RGB object crops. Four stride-two convolution blocks use channels 3/32/64/128/128 with BatchNorm and ReLU, followed by adaptive pooling and a normalized 128-dimensional projection. Predicted color, shape, material and size classes accompany the crop representation. The Full head has two pre-norm Transformer layers, four attention heads, hidden dimension 128, feedforward dimension 512, dropout .1, candidate embeddings of size 64 and attribute embeddings of size 32. Position/type embeddings distinguish the ordered object/relation tokens. Object IDs index selected cache entries rather than entering the answer head as target-revealing inputs.

The encoder's locked training and selection manifests contain 23,662/4,467 images and 156,724/29,428 crops. Their intersections with the discovery and extension panels are empty. The head-training and development images also have no recorded encoder-membership overlap. This establishes documented exposure, not absence of undocumented experiments or independence from unknown external pretraining.

The controlled four-selection benchmark, called SI4 in the experiment records, has 512 training, 128 development and 128 discovery images. Training/development cue–target agreement is .875; discovery is .50. The 512-image extension has .2344 agreement and is deliberately a cue-shift stress test. Initial heads use AdamW .0015, weight decay .0001, batch 256, at most 50 epochs and patience eight. Development accuracy, then NLL, then earliest epoch selects Initial. The fixed visual seeds are 17, 29,43. A high cue-correlated development score is not an out-of-distribution task-capability guarantee.

B.2 COMPLETE VLM INTERFACE AND ADAPTATION

The Four interface is a 408×408 labeled montage with four 192-pixel tiles and fixed gaps. Each tile is constructed from the same centered crop source. The question explicitly compares the centered objects in tiles 1 and 2; the other two are legal context. The processor receives the full montage. Crops can include other scene objects, and visual contextualization can mix tile information. Input-position support is therefore distinct from an isolated object-attribute claim.

The exact prompt for the size question is:

> The image is a labeled montage with 4 tiles. Compare only the centered object in tile 1 and the centered object in tile 2. Other tiles are legal context and do not change which two objects are compared. Do the centered objects in tile 1 and tile 2 have the same physical size category? Answer Yes or No only.

For material, the comparison asks whether the two centered objects are made of the same material. Native Pair changes the declared tile count and supplies only the task crops; processor resizing can also differ. It is an access control, not an identical internal-state control.

Qwen3-VL-30B-A3B-Instruct is a complete mixture-of-experts model. Its initial task adaptation uses language-attention LoRA with rank 16, alpha 32 and dropout .05, fitted on 512 original Four examples. Eight epochs yield 4,096 presentations and 1,024 updates; the adapter contains 13,369,344 trainable parameters. The final adapter obtains Four 127/128 and Pair 125/128 on its development check, with material 64/64 and size 63/64 in Four. Adaptation is not CF repair. The model and adapter remain fixed during the later interface, exchange and small-correction experiments. The exact target modules are the language model's `self_attn.{q_proj,k_proj,v_proj,o_proj}` in every layer; visual weights are not adapted. AdamW uses learning rate $2\,10^{-4}$, zero weight decay, microbatch one, four-step accumulation, gradient clipping one, cosine scheduling and 3% warmup. Seed 358 and the final eighth epoch are fixed before evaluation; only answer tokens contribute to

the language-model loss. The pre-output capability gate requires at least 95% Four accuracy overall, 90% per attribute, 99% valid answers and a Pair–Four accuracy difference no greater than five percentage points. The same four-tile thresholds apply to the native 30B and 38B checks.

Greedy generation uses at most 16 new tokens, BF16 and no quantization. Only a complete Yes/No answer with optional punctuation is valid. Earlier observational audits count invalid responses as errors and separate support-output classes; the later mechanism pilots instead stop on invalid generation or scoring/generation disagreement. No such failure occurred in their completed evaluations. The recorded environment uses PyTorch 2.11.0, Transformers 5.12.1, PEFT 0.19.1 and NumPy 1.26.4 on an H100.

B.3 Native-frozen model checks

Table 8: Native-frozen small-VLM exploratory audit. Thirty-two opened scenes, 836 routes, no CF tuning. Accuracy is image-equal. All fail the pre-output Pair development gate of at least 14/16 correct. These are not task-capable external confirmations or model rankings.

Model	Pair dev	Four acc.	$\mathcal{U}^\star$ miss/pass	$\mathcal{D}$ miss/pass
Qwen3.5-9B	12/16	61.46	2/6	5/6
InternVL3-8B	13/16	67.19	2/7	4/7
LLaVA-OneVision-7B	12/16	57.51	1/4	3/4

Table 8 retains the original exploratory outcomes (Qwen Team, 2026; Zhu et al., 2025; Li et al., 2024). Their original-fiber local-miss counts are zero, so these broader selected-prefix results do not replicate the exact endpoint example. Native InternVL3-38B answers 108/128 Four and 117/128 Pair development queries correctly and fails its specified capability gate. Its 24-image exploratory restriction was adopted after partial runtime outputs, not prospectively frozen as a confirmation panel. It gives 80.98% full-domain accuracy, zero selected-prefix misses and six full-domain misses among 11 local passes.

The adapted Qwen's earlier 116-image audit gives 96.64% full-domain accuracy and 7/103 selected-prefix misses. Although adapter-fit memberships are disjoint,104/116 images had appeared in earlier project evaluations, including all seven residual images. We retain this panel as exploratory and use later registry-disjoint panels for fresh-image evidence.

C Domain Structure and Support Consistency

This section separates two reasons an audit can be uninformative. Geometry may prevent the needed comparison, even for a simple rule. Alternatively, the queried set may contain no repeated support value whose predictions can be compared. The first argument concerns legal paths; the second concerns what observed answers can rule out.

C.1 The diameter bound and its conditions

For the original relation rule, a surprising-looking two-edit failure has a simple geometric explanation: every legal suffix is reachable within two edits. Here is the argument and its scope. For a fixed task prefix, let $\mathcal{A}$ be the legal third-position objects. A fixed spatial direction can be represented by a scalar projection ℓ, with the last relation requiring $\ell(w) - \ell(u) > \varepsilon$. If $\ell(u) \le \ell(u')$ and (u', w') is legal, then (u, w') is legal. Hence

$$(u, w) \longrightarrow (u, w') \longrightarrow (u', w')$$

joins any two legal vertices in at most two one-position edits (reverse the ordering argument if needed). The result assumes the enumerated spatial relation and no extra joint constraint excluding the intermediate input. It is not implied merely by having two editable positions.

Actual geometry is checked from stored scene relationships. On the 220-image panel, Relations and All-distinct fixed-prefix fibers are nested, connected and have diameter at most two. A supplemental root-conditioned bounded restriction produces 164 nonnested fibers,81 disconnected fibers and 104

with a finite component diameter above two; counts overlap. This artificial restriction tests structural scope, not new natural-image generality. Disconnected graphs have no finite global diameter.

Equation 4 holds without nesting or connectivity, because it directly enumerates the legal endpoint sets accessible from the root. When the original fiber has distance-two vertices, every nonlocal error is necessarily at distance two. That location is a consequence of geometry and selection, not another independent mechanism discovery.

C.2 Compatibility, gap and non-identifiability

The support check asks whether equal proposed inputs ever yield different predictions. For finite $\mathcal{Q}$, Equation 3 is necessary because a function cannot return two answers for one support value. It is sufficient because the common observed answer can be assigned to that value. Grouping by the selected positions takes linear expected time in the number of queries using a hash map.

The support gap measures how far the observations are from satisfying this condition. Let $n_{z,y}$ count queried inputs with support value z and prediction y in a nonempty query set $\mathcal{Q}$. An empty query set is marked unevaluated, not assigned a zero gap. The smallest fraction of observations that must be removed to make every group consistent is

$$\epsilon_J^{\mathcal{Q}} = \frac{\sum_z \left(\sum_y n_{z,y} - \max_y n_{z,y} \right)}{|\mathcal{Q}|}. \tag{7}$$

To see why, retain the most frequent answer in each group and remove the others. No smaller deletion can make that group constant. This is an approximate functional-dependency measure, not an accuracy measure: a group that always gives the wrong answer still has zero gap.

For a binary task-prefix group with n_z inputs, e_z errors and one common target,

$$\epsilon_{01}^{\mathcal{Q}} = \frac{\sum_z \min(e_z, n_z - e_z)}{|\mathcal{Q}|} \leq \mathrm{err}_{\mathcal{Q}}(\hat{f}, T).$$

A consistently wrong group contributes no gap. If S contains only new, mutually distinct J projections absent from $\mathcal{Q}$, then

$$\epsilon_J^{\mathcal{Q} \cup S} = \frac{|\mathcal{Q}|}{|\mathcal{Q}| + |S|} \epsilon_J^{\mathcal{Q}}.$$

These statements characterize the measured criterion; they are not new general dependence theorems. They connect classical approximate dependencies and constrained tests to the actual visual domain (Mandros et al., 2017; Jin & Tsuchiya, 2020).

The controlled 512-image extension illustrates two different dilution effects. After 160 CF updates, 68/1,536 model–image records have a nonzero task-support gap. Within those same records, the mean gap is 24.95% when normalized by the original-pair set $\mathcal{F}$, but 15.69% using the larger $\mathcal{U}$. Added singleton task pairs increase the denominator without adding contradicting pairs. Averaging over the other zero-gap records then reduces the overall mean to 0.695%. These are distinct operations: enlarging each record's input set and averaging over more records. Completing task pairs addresses the first issue by adding repeated comparisons on 503/512 images. The total inputs per head are 6,418 in $\mathcal{U}$, 30,595 in $\mathcal{U}^\star$ and 104,055 in $\mathcal{D}$.

Randomization also needs room to change the behavior being tested. A control that preserves both target and endpoint cue, called Target+cue, can preserve every error of an exact endpoint rule. It therefore cannot test whether that shortcut exists. When conditioning leaves no possible rearrangement, the check has no freedom to distinguish alternatives and must be reported that way. Likewise, an unchanged Yes/No answer can hide a probability change; the hard-answer gap does not measure that effect. Separate probability analyses remain in the historical evidence archive.

Under a task-local pass, each prefix in $\mathcal{U}^\star$ has a correct queried anchor. Any incorrect completion shares the target but not the prediction, providing a task-support counterexample. No analogous implication holds for a previously untested prefix in $\mathcal{D} \setminus \mathcal{U}^\star$. Separately rejecting $\mathcal{H}_{\{0,1\}}$ and $\mathcal{H}_{\{3\}}$ does not reject $\mathcal{H}_{\{0,1,3\}}$; the latter allows joint dependence, not merely the union of the former two hypothesis sets.

C.3 What must be queried to certify a support claim?

The consistency test has an important asymmetry: two inputs can reject a support explanation, whereas certifying it without assumptions on unqueried behavior can require an entire group. The following finite-domain statement makes the audit's unresolved output precise. It is an application of functional dependence and indistinguishable black-box completions, not a new general learning bound.

Proposition 2 (Universal compatibility certificate)**.** *Fix a finite legal domain $\mathcal{D}$, support projection r_J, and at least two possible output labels. Let $\mathcal{Q} \subseteq \mathcal{D}$ have observed predictions consistent with $\mathcal{H}_J$. Every deterministic extension of those predictions to $\mathcal{D}$ belongs to $\mathcal{H}_J$ if and only if every member of every nonsingleton J-group in $\mathcal{D}$ has been queried.*

Proof. If all nonsingleton groups are observed, consistency fixes a common answer for each such group. A singleton imposes no equality constraint, so any extension is compatible. Conversely, let a nonsingleton group contain an unqueried input u. If another member has been queried, assign u a different label. If no member has been queried, assign two members different labels. Either assignment extends the observations but violates $\mathcal{H}_J$. □

This is a universal statement about all black-box extensions, not a claim that the actual model violates support whenever a query is missing. It also distinguishes the usefulness of singleton inputs: they can find an accuracy error, but cannot contradict support by themselves. The black box is indexed by routes. Known identities of actual model inputs or certified invariances can constrain the extension class and allow additional reuse; the lower bounds below do not charge another call for an input already known to be identical.

Proposition 3 (Completion after a local pass)**.** *Suppose the task target is a deterministic function of the prefix and all inputs in $\mathcal{L}_{\text{task}}$ are correctly answered. On $\mathcal{U}^\star$, task-only support is equivalent to perfect accuracy. In the worst case, any zero-error black-box procedure certifying either property needs all $M = |\mathcal{U}^\star \setminus \mathcal{L}_{\text{task}}|$ additional queries. If $k > 0$ of these M inputs are wrong, a uniformly sampled subset of $0 \le b \le M$ inputs, without replacement, misses them with probability*

$$\frac{\binom{M-k}{b}}{\binom{M}{b}} \;\le\; \exp(-bk/M),$$

where the numerator is zero for $b > M - k$.

Proof. Each prefix in $\mathcal{U}^\star$ has a correctly answered local anchor. Constancy within its group therefore implies correctness throughout that group; correctness implies constancy because its target is fixed. Leaving any additional input unqueried permits an extension differing from the target only there. The extension produces the same transcript as the correct rule but violates both properties. Thus an always-correct certificate cannot stop early on the all-correct transcript. For uniform sampling, $\binom{M-k}{b}$ subsets avoid all errors out of $\binom{M}{b}$ possible subsets. Multiplying the successive avoidance probabilities gives the exponential bound. □

The sampling formula assumes uniform queries and a fixed error set; it does not apply to a confidence-ranked adaptive policy without further assumptions. Nor does the lower bound make completion optimal for finding the first error. It specifies the information needed for an exhaustive certificate on the selected prefixes. Neither proposition assumes a connected edit graph or a particular spatial relation. Only the finite domain, support projection and, for the second proposition, a known correct anchor per selected prefix are required.

C.4 Sharp bounds when some comparisons remain unqueried

The unqueried scope can also be expressed as an exact identification interval for the binary support gap. Let the full domain contain $N > 0$ inputs, with N_z in support group z. Among queried inputs, let $n_{z,0}$ and $n_{z,1}$ count the two answers and let $m_z = n_{z,0} + n_{z,1} \le N_z$. Across all deterministic

binary extensions of these observations, the smallest and largest possible full-domain gaps are

$$\begin{aligned}\epsilon_{\min} &= \frac{1}{N}\sum_{z}\min(n_{z,0}, n_{z,1}),\\ \epsilon_{\max} &= \frac{1}{N}\sum_{z}\left[N_z - \max(n_{z,0}, n_{z,1}, \lceil N_z/2\rceil)\right].\end{aligned} \tag{8}$$

Both endpoints are attainable. To minimize the minority count in a group, assign all unqueried answers to an observed majority. To maximize it, balance the final labels as far as the observed counts allow. The smallest achievable majority is $\max(n_{z,0}, n_{z,1}, \lceil N_z/2\rceil)$. Groups can be completed independently, proving sharpness after summation.

This is a logical range, not a confidence interval or a probability of shortcut dependence. It assumes valid deterministic binary answers; invalid outputs remain a separate audit state. A singleton contributes zero to both bounds, even if unqueried. With consistent observations, $\epsilon_{\max} = 0$ exactly when all nonsingleton groups have been exhausted, recovering the first certificate. With contradictory observations, the lower bound remains positive whatever the unqueried outputs are. Thus the same returned audit record distinguishes observed violations, undetermined violations and purely denominator-changing additions.

C.5 A CONSTRAINT-STRENGTH COMPARISON NEEDS A FIXED REFERENCE

Hold the predictor, target and reference set $\mathcal{S}$ fixed. For query sets $\mathcal{Q}_0 \subseteq \mathcal{Q}_1$, define the blind event as correctness on $\mathcal{Q}_i$ together with an error in $\mathcal{S}$. The blind-event set for $\mathcal{Q}_1$ is a subset of that for $\mathcal{Q}_0$, because passage of the larger query set implies passage of the smaller one. This deterministic fact does not predict how many events disappear. It also does not compare two changing denominators or two different reference domains. Our relation-removal experiment holds the original completion set fixed and measures which of its misses become visible after adding formerly excluded one-edit inputs.

C.6 WHY SEMANTIC CONTROL DOES NOT IMPLY NATURAL-ERROR ATTRIBUTION

A minimal counterexample clarifies the separate empirical tests. Let a representation have binary semantic and context components (x, c), true target $T = x$, and readout $\hat{f}(x, c) = x \oplus c$. On a reference panel with $c = 0$, replacing x by a donor value gives the correct composed target with probability one. Thus semantic exchange can be perfectly validated on that panel. Natural correct and incorrect runs $(x, 0)$ and $(x, 1)$ have identical semantic components, however. Transferring either semantic donor into the same third receiver has exactly the same effect. Replacing the full state can distinguish the donors. The example demonstrates non-implication, not an identified XOR computation in the VLM. It motivates the fixed-third-receiver control and the separate comparison of intervention-informed and directly learned corrections. Neither semantic control nor successful restoration logically guarantees an incremental repair benefit.

C.7 AUDIT IMPLEMENTATION

Table 9 gives the procedure. Linear grouping cost excludes legal-domain enumeration and model inference.

D CONTROLLED IMPROVEMENT AND INDEPENDENT CONFIRMATION

These experiments answer three different questions. The matched continuation compares training recipes on one fixed panel. The two-task study tests repair on independent images. Fresh-head and isolated-rendering controls then test whether audit blind spots remain after improvement or removal of crop context. Their results are not pooled into one accuracy estimate.

D.1 MATCHED CONTINUATION

Original CE, Balanced CE and CF share initialization, optimizer reset, original-image order and effective batch size. AdamW uses .0005, weight decay .0001 and clipping 1. Two batches per epoch

Table 9: Finite-domain audit pseudocode. A budget can terminate at any query; compatibility is then restricted to the queried set. Support rejection does not require task labels.

1	Enumerate legal routes from the fixed scene and relation program.
2	Construct the root's legal task-local set; freeze the selected prefix set.
3	Query chosen routes or reuse an identical-inference cache; deduplicate.
4	For each r, index prediction $\hat{f}(r)$ by projection r_J.
5	Return any pair with equal projection and different prediction.
6	Count unqueried inputs; report the observed gap and, for valid binary outputs, the extension bounds in Equation 8.
7	Mark the group contradicted, exhausted without contradiction, or unresolved.
8	After an update, reuse the same domain and compare repaired, persistent and new error IDs.

each contain 256 originals and 256 counterparts. Balanced CE weights cue-agree/ conflict originals by 4/7 and 4; CF averages original and legal alternative losses. Budgets 40/80/160 are frozen and every trajectory runs to 160. Alternatives come only from training images. Table 10 retains every fixed endpoint.

Table 10: All matched controlled-head budgets. Accuracies are image-equal percentages on Extension512; counts pool three fixed heads. BC counts correct target-changing pairs. Conditional misses use task-local passage and an original-fiber error.

Training / updates	$\mathcal{F}$ acc.	$\mathcal{U}$ acc.	BC/483	Miss/pass
Original40	46.85	50.67	50	23/68
Original80	50.84	54.05	82	43/107
Original160	52.16	54.93	86	40/106
Balanced40	52.45	54.26	75	35/112
Balanced80	54.28	56.70	98	49/146
Balanced160	55.24	57.52	92	33/149
CF40	62.35	64.57	156	49/357
CF80	91.59	91.79	393	37/1048
CF160	97.99	98.19	459	11/1416

The stronger full-input reference changes architecture, augmentation and development selection jointly. It attains 99.79% fiber accuracy and no anchor-domain local misses, but still has 13 full-domain misses when the reference scope expands. This is not deterioration of an unchanged model or a pure architectural ablation.

D.2 Independent two-task repair

Endpoint asks the final object's attribute; Source–interior compares the first two selected objects. Both use supplied routes, frozen crops and predicted classes, without updating a router or encoder. Each task uses 1,200 training images. The six training arms are fixed before evaluating 640 new images,320 per task, disjoint across tasks and recorded memberships. Accuracy uses 160 cue-conflict images/task; nuisance total variation (TV) and target-changing BC use all 320.

Table 11: All independent two-task repair arms. Accuracy and BC are fractions; TV measures nuisance probability variation. The original primary comparison was Adaptive ICRA minus Before, not a retrospectively selected ordinary-CE superiority claim.

	Endpoint			Source–interior		
Method	Acc.	TV	BC	Acc.	TV	BC
Before	.5542	.2967	.5573	.9021	.0915	.8958
Original CE	.5125	.3015	.5156	.9104	.0584	.9125
Random legal CE	.9917	.0028	.9875	.9917	.0018	.9906
Stratified CE	.9917	.0018	.9875	.9917	.0015	.9885
Adaptive CE	.9896	.0025	.9885	.9896	.0024	.9885
Adaptive ICRA	.9896	.0019	.9885	.9917	.0019	.9906

Each repaired arm starts at the same task/seed-specific head. AdamW uses .0005, decay .0001, clip 1, 16 epochs and base batch 128. Original and two counterpart branches have weights .5/.25/.25; each arm receives 160 updates and 57,600 forward examples, with dropout disabled equally. I-CRA adds invariant-distribution JS and target-aligned margin objectives, both weight 1 with target-probability margin .2. All arms pay the same audit-forward allowance. Table 11 establishes ordinary legal-CE gains but no superiority of ICRA over that simpler alternative. The original ICRA gains are 43.54[37.92,49.17] and 8.96[6.46,11.67] percentage points. Intervals resample images 10,000 times with seed 22301 and retain the fixed visual seeds.

D.3 Fresh heads and isolated rendering

The 220-image confirmation uses heads 101/103/107 sharing encoder 17. Development accuracy/NLL selects Initial; exactly 160 CF updates define the improved state. Relations and All-distinct use matched roots. This is a before/after confirmation, not a new training-recipe comparison. Fixed-head image-cluster resampling yields the 17.16% selected-prefix risk interval in the main text. Its head-specific risks range from 4.31% to 29.03%, so pooled counts do not establish uniform behavior across initializations.

The isolated-rendering control replaces crops with single-object rerenders, retrains the CNN and predicts new attributes. Geometry, fit splits, CF pairs and budgets remain fixed. Evaluation uses new render variants on 220 already opened geometries, not 220 unseen scenes. The shared finite render bank and nonuniform capability are explicit constraints. Table 12 retains every head.

Table 12: All isolated-rendering heads, Relations legality. The development gate is at least 90% CF160 accuracy. No head is excluded. Joint witnesses require a same-prefix contradiction and correct task attributes/Lookup at a wrong input.

Head	Dev gate	CF160 $\mathcal{D}$ acc.	$\mathcal{U}^\star$ miss/pass	$\mathcal{D}$ miss/pass	Joint/miss
101	Fail	77.63	41/74	59/74	57/59
103	Pass	88.28	46/122	87/122	80/87
107	Pass	99.48	6/217	29/217	29/29

A separate 372-image cross-prefix study starts after CF160. Target-changing legal-pair training resolves all 13 observed majority-vote residual images. However, disagreement ranking and a post hoc matched target-NLL control select almost identical examples and have identical reported outcomes. The initial residual support is below the protocol minimum 16; target change is not established as necessary or the ranking uniquely useful. This evidence remains bounded and does not replace the independent two-task result.

E Queries Resolve Different Questions

An audit may aim to find any error quickly or to finish checking a particular support explanation. Those goals need not favor the same query order. The comparisons below charge the same query cost and report both outcomes, rather than treating one as a proxy for the other. The query studies begin from the same fully observed $\mathcal{U}$ and use fixed geometry and common hashes, never unqueried predictions. Closure-first prioritizes missing suffixes of selected prefixes; Uniform ranks all remaining $\mathcal{D}$ inputs; Prefix-balanced chooses the least-queried prefix. All queries are charged, including those after finding an error. Table 13 separates their outcomes.

Within a fixed set of task pairs, completion resolves 66.04% of checks, compared with 56.54% for hash order at the same mean 28.54 total calls. A check is resolved either by finding a contradiction or by exhausting all its legal inputs without one. It is not necessarily correct. Thus an order can finish more support checks yet find fewer errors overall. Exact-closure-cost and nonexhausted-pool controls remain in the historical source tables; no globally optimal query policy is established.

Auxiliary task-local TV prioritization does not establish an advantage over original confidence or local minimum margin on the main confirmation budgets. Fixed-prefix TV is a post hoc sensitivity analysis. Binary confidence, entropy and Gini have equivalent rankings for a single prediction and

Table 13: Equal-cost expansion on the opened Confirmation220 Relations panel after CF160. Error/rejection recalls concern events absent in $\mathcal{U}$ and present in $\mathcal{D}$; their denominators are 176/231 records. Contradictions count newly contradicted selected prefixes per record. Results average three heads and 32 common orders.

16-call allowance	Error recall	Support rejection recall	Selected contradictions
Closure-first	42.21	52.34	.324
Uniform	60.92	53.86	.182
Prefix-balanced	64.90	39.33	.100

are not counted as three independent baselines. Total query cost includes the local audit as well as upgrades. These limits constrain practical recommendations without becoming a separate main contribution.

F Matched Complete-VLM Continuation

This study tests example selection, not a new repair architecture. It asks whether training on support counterexamples repairs known residual cases better than ordinary choices under the same budget. The fixed initial error cohort is retained even when a repaired model no longer passes the local audit on those images. The 128-image continuation panel is selected from 4,492 remaining validation scenes outside a searched registry of 34,505 image names. Scenes, attribute, relation program and root are hash-ordered before outputs, with 32 images per attribute/root-target cell. Each has 4–64 full-domain inputs, a nontrivial fiber and local neighborhood. The result has 3,173 inputs in $\mathcal{D}$ and 1,867 in $\mathcal{U}^\star$. It is newly constructed task sampling on unused scenes rather than another sample of the inherited question manifest.

All arms use the same common anchors and 1,268-candidate pool, charged equally before training. Anchor CE repeats anchors; Local CF and Uniform CF pair them with their fixed alternatives; Max-NLL selects the largest target-token NLL over the full vocabulary. Witness/completion uses a same-prefix prediction contradiction from a correct anchor when available, otherwise a hash-selected completion. Only 2/128 images provide the former. Selection recipes differ in target balance and difficulty, so this is not an isolated edit-distance effect or a reproduction of the full JTT procedure (Liu et al., 2021).

Each arm receives 256 examples, AdamW reset, learning rate 5×10^{-5}, weight decay 0, microbatch 1, accumulation 4 and clip 1. Exactly 64 updates define the final checkpoint. Only the existing task-adapter parameters update; answer tokens and end-of-turn are supervised after exact prompt-prefix masking. Seeds 17/29/43 share the adapted initialization. Table 14 reports all recipes on the same cohort.

Table 14: All continuation recipes on the fresh 128 panel. Accuracy averages within images and over three continuation seeds. Remaining errors count the fixed five initial residual images, including those that become local failures.

Recipe	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ acc.	Remaining/5, by seed	Target-change BC
Initial	97.34	97.43	5	94.29
Anchor CE	97.31	97.53	3/3/2	94.36
Local CF	97.19	97.36	3/2/4	94.58
Uniform CF	97.01	96.98	3/3/3	94.15
Max-NLL	97.85	97.99	3/4/4	96.30
Witness/completion	97.45	97.60	3/3/3	94.89

The prespecified contrasts are witness/completion minus Uniform and Max-NLL on the fixed initial residual cohort. The first is 0.00[-26.67,26.67] percentage points; the second is$-13.33[-26.67, 0.00]$. Max-NLL's secondary full-domain gain is 0.51[-0.37,1.47] points. These image-bootstrap intervals retain the three fixed continuation seeds and do not support a precise ranking.

Table 15: **Four tests separate internal control from error explanation.** Populations differ across rows. Full-slot is higher-capacity; Semantic and answer-preserve are rank-matched. Appendix G gives uncertainty and sampling.

Test	Decisive observation
Task-state replacement	Restores 8/10 new-image errors; normal-control damage: 1/20 directions.
Attribute readability versus use	Probe-axis changes 126/128 decoded attributes, but 0/128 expected answers; Full changes 121/128.
Held-out semantic exchange	Semantic agreement: 95.14% change/retain, 97.22% unseen pairs; answer-preserve change: 0.69%.
Natural-error donor, fixed receiver	Correct/error-context donors give different hard answers in 0/28 Semantic tests and 7/28 Full tests.

The main inference uses batch 8. A later replay of all 174 inputs in the five residual images, across 16 model states, changes 15/2,784 answers at batch 1. All Initial local sets remain correct, but one Initial residual disappears. The original cohort and batch 8 analysis are retained; batch 1 witness-versus-Uniform remaining-risk difference is$-6.67[-26.67, 13.33]$ points. The representative witness in Figure 2 survives both. Subsequent experiments explicitly fix batch 1.

Candidate scoring uses 1,268 responses; the 15 continuations use 3,840 training presentations. Main inference uses 50,768 generated responses, Pair/root replay adds 15,248, and complete cohort replay adds 2,784. Together with scoring this gives 70,068 response evaluations, in addition to training. Later mechanism reports count scoring forward and actual generation as separate top-level calls; those units must not be pooled with this response count as if they were identical FLOPs.

G Internal Interventions and Semantic Alignment

An internal replacement copies selected states from one run into another. The supplying run is the donor; the receiving run keeps all unselected states. The experiments progress from testing where replacement has an effect to testing what that effect means. Processor checks, failed interventions and successful interventions are all retained because they constrain the interpretation differently.

G.1 Interface validation and a stopped layer hypothesis

Coordinate-coded inputs validate 676 raw patch positions and their 13-by-13 merged grid. Each tile contributes 25 strict-interior positions;69 positions cross boundaries, gaps or outer edges. We avoid inferring ownership by taking the square root of a token count. Actual-generation controls require self-patching to preserve predictions and full visual replacement to reproduce donor output.

An initial directional-attention pilot patches the changed tile's visual K/V only at the final question-position receiver in zero-based decoder layers 4, 8, 12, 20, 28,36. No layer restores any of four development errors. The best median directional effect is .0625 logits, below the frozen .5 threshold. This specific branch stops; it does not rule out distributed or upstream causes.

The visual-interface branch instead freezes both task slots and all four visual channels after development. On 256 new images, original full-domain accuracy is 96.02%, with 10 selected-prefix misses among 227 local passes. All ten selected pairs have target No, comprising eight size and two material examples. Their 8/10 forward restorations and 1/20 normal-direction damage rates cannot be generalized to Yes-target errors or a broad population without new evidence.

The median directional margin effect is 4.03125 logits, with descriptive interval[3.03125,7.875]. Its equal-position-count task-minus-suffix contrast has median 1.03125 and interval [.0625,5.5]. Equal counts do not mean equal semantic content or equal perturbation norm. A supplemental norm attenuation happens to use coefficient 1 in every case, so it is not an independent amplitude-robustness result.

G.2 COMPOSITION AND NATURAL-ERROR TRANSFER

Attribute-composition tests ask whether the patched answer follows the attributes now represented in the two task slots. Natural-error transfer instead asks whether states from a wrong run carry a reproducible error effect. A successful composition test does not, by itself, answer the second question. On 32 new-image correct/correct semantic pairs, single-slot full replacement matches the attribute-comparison target on 120/128 interventions; double-slot replacement matches on 62/64. The single-slot rule differs from copying the donor's original answer. However, this test does not exclude every distributed answer-control explanation.

The error-context study fixes 18 development error pairs and 19 fresh-image error pairs. Of these, 12 and 16 respectively admit a third receiving run for a matched-donor comparison. We replace the first task group, second task group or remaining positions, alone and in combination. Replacing both task groups restores 12/18 development and 12/19 fresh errors. Replacing only the remaining positions restores 12/18 and 13/19. More than one internal replacement can therefore recover the same type of failure (Table 16).

The predefined task-state signature occurs in 1/12 development and 3/16 fresh third-receiver cases. The rest-state signature occurs in neither group, giving 0/12 and 0/16. Neither meets the replication gate. These tests distinguish particular transfer patterns; they do not exhaust all possible perception, binding or comparison failures.

Table 16: Full-state source decomposition on error-selected pairs. Bits indicate whether task0/task1/rest comes from the correct run. The all-zero/all-one rows are replay controls, not discoveries.

Restoration	Development correct/18	Fresh correct/19
000	0	0
100	10	10
010	10	6
110	12	12
001	12	13
101	15	16
011	15	16
111	18	19

G.3 PROBE-AXIS EXPERIMENT

The probe is a separate attribute reader, not the VLM's own answer head. This experiment changes the direction detected by that reader and checks whether the original model follows the predicted attribute change. For each tile, the 25 internal positions are averaged separately over four 2048-dimensional channels and concatenated to 8192 features. Each attribute uses a shared linear ridge readout fitted on 48 images and validated on 24 separate images;32 further images calibrate interventions and 37 error images assess readability. These are disjoint roles within previously opened material. Training uses standardized features, variance floor10^{-3}, fixed squared-error loss with L2 weight 1 and threshold 0. A scene/object-hash pseudolabel control scores about chance.

Axis intervention projects the donor/receiver mean difference onto the normalized probe direction and broadcasts that component to the 25 positions. Random uses an orthogonal unit direction with the same requested magnitude; Complement subtracts Axis from the full token difference. Axis changes the probe class 126/128 times but causes 0/128 expected single-slot answer changes. Full and Complement both succeed 121/128 and have identical hard answers across 192 conditions. This comparison concerns one fixed broadcast direction, not all representations detectable by probes.

G.4 LEARNED SEMANTIC MAPS

Unlike a probe fitted only to read an attribute, the semantic map is trained on the VLM's answer after replacement. The answer-preserve control uses the same replacements but is trained to keep the receiving run's original answer. Separate material/size maps have four rank 4 orthonormal channel bases, shared over positions and task slots. Each has 32,768 trainable scalars. Map fitting uses

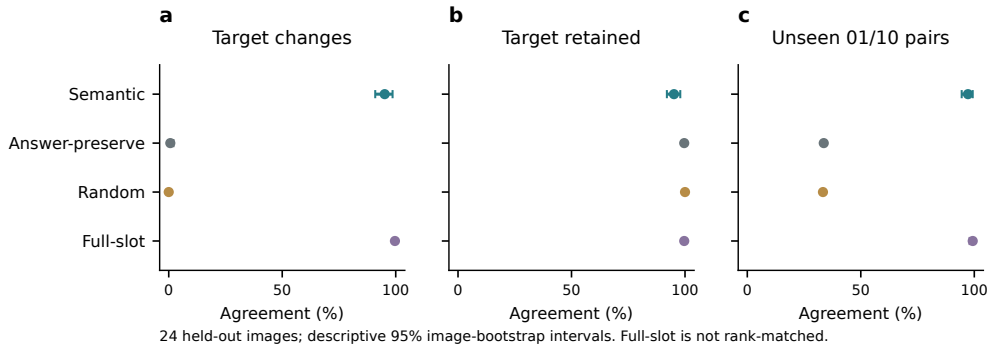


Figure 5: **A learned exchange can change answers in the predicted way.** Agreement is with the answer implied by the exchanged attributes, not with the donor's answer. The panels separate target-changing, target-preserving and unseen 01/10 pairs. All use the same 24 held-out images and 288 interventions per displayed condition. Bars show descriptive 95% image-bootstrap intervals. Semantic and answer-preserve maps are matched; Random is untrained and Full-slot replaces more dimensions. These are intervention tests, not deployment repair rates.

48 images,24 per attribute; evaluation uses 24 other images,12 per attribute. All come from the previously opened 576-image pool and exclude the 213 probe-study images. The 37 natural error pairs do not enter map optimization.

Correct/correct route pairs are selected by fixed hashes. Training uses 00/11 attribute-change types; evaluation includes 00/11/01/10, both directions, and task0/task1/both exchanges. Semantic and answer-preserve objectives share initialization, data and update order. AdamW uses .02, zero decay, epsilon10^{-8} and clip 1 for two fixed epochs,576 steps per fitted map. Only the map receives gradients; the VLM, adapter and dropout remain frozen/evaluation-mode. A non-reentrant decoder checkpoint path is validated before fitting. The final step is retained without evaluation-based selection. Table 17 separates semantic agreement from natural-error transfer.

Table 17: Exchange agreement and selected natural-error transfers. Change/retain/unseen are percentages on 24 held-out images. Correct-donor and reverse-damage columns use 37 error-selected images, not the same evaluation population or deployable repair rate.

Map	Change	Retain	Unseen	Correct donor	Reverse damage
Semantic	95.14	95.14	97.22	23/37	12/37
Answer-preserve	0.69	99.65	33.68	5/37	1/37
Random	0.00	100.00	33.33	4/37	0/37
Full-slot	99.65	99.65	99.31	24/37	12/37

For fixed receiver C, semantic correct/incorrect-context donors produce 23 pairs of correct answers and five pairs of incorrect answers, with no hard-answer disagreement. Full-slot replacement has 15 both-correct, seven only-correct-donor-correct, and six both-incorrect cases. The semantic target-margin difference averages .237 logits with descriptive interval[.049,.438]. These data limit error-transfer fidelity while leaving the learned semantic intervention result intact. Broader criteria for preserving non-target properties are motivated by RAVEL (Huang et al., 2024); our interchange benchmark is not a full RAVEL-style disentanglement evaluation.

H Historical Reference-based Corrections

These frozen rules test whether a label-free donor is itself sufficient for repair. They precede the final write-space study and are not retuned. The complete-record companion retains every rule, interval and cost.

An isolation reference blanks suffix tiles while keeping the task tiles, labels and canvas. On 6,431 legal inputs from 256 images, half-strength mixing repairs 38 errors and introduces 32. Image-equal accuracy nevertheless falls by 0.028 points. Only two of ten original completion-miss images become fully correct in that scope. Seven errors arise even where both original and reference answers were correct. Reference correctness alone therefore does not ensure successful internal transfer.

A separate 128-image legal-reference pilot has no eligible completion misses; that endpoint is unevaluable, not zero-percent repair. Full task-state injection reaches 98.41% full-domain accuracy, below direct reference use at 98.55%, and fails its incremental gate.

The later semantic-mean rule uses two hash-selected legal contexts per task pair, repeating the unique input for singleton pairs. No prediction, label or confidence selects a donor. It acts on 50 task positions and is compared with AP, random, norm-matched random, complementary, full-state, direct-reference and voting rules on 128 images (2,923 inputs, 818 prefixes). The semantic rule repairs ten errors and introduces nine. Its accuracy gain is 0.082 [-0.451,0.647] points; regression is 0.611%, above the prespecified 0.5% limit. It repairs one of two local blind spots, retains one and creates one. The gate fails despite sufficient error-image support. No map, reference selector or VLM is fitted in this experiment. The complete record includes all alternatives and 42,458 scoring/generation calls. These results motivate donor-free correction, not a claim that reference quality is the unique cause of failure.

I Historical Donor-free Pilot

The first donor-free correction uses the earlier exchange maps, not the separately constructed final write spaces. In stream c, it writes $h'_c = h_c + Q_c(A_c^\top x_c + b_c)$, where $x_c = h_c/\sqrt{\mathrm{mean}(h_c^2) + 10^{-6}}$. Four fixed rank-four bases give 32,784 trainable coefficients and biases per attribute; learning the bases too doubles that capacity. Zero initialization reproduces the original states.

Training uses 512 inputs from 64 scenes, with two epochs and 512 updates per attribute/arm. AdamW uses learning rate 10^{-3}, zero decay, epsilon 10^{-8} and clipping one. The binary CE adds 0.01 times the mean squared update norm. All arms share data, order and optimization; only 50 task positions are written, with BF16 writeback. Deployment needs no donor or label. Evaluation reuses the opened 128-image reference-mean panel.

Table 18: All historical donor-free pilot arms. Accuracy and regression are image-equal percentages; repair/new are input counts. This single-fit development comparison is not the final confirmation. The free-space arm has twice the capacity; Probe has extra supervision.

Correction	D accuracy	Repair / new	All-correct	Regression
Original	96.92	0 / 0	107/128	0.00
Semantic	97.18	71 / 70	109/128	2.48
Answer-preserve	98.81	68 / 6	118/128	0.32
Random	97.56	18 / 3	111/128	0.11
Free subspace	98.25	47 / 35	113/128	0.85
Probe + equality	98.58	74 / 54	121/128	1.26

The primary Semantic arm repairs 71 full-domain inputs but introduces 70. In completion it repairs 24 and introduces 56; outside completion it repairs 47 and introduces 14. Its mean regression, accuracy gain and completion bad-prefix change fail the prespecified expansion rule, so the conditional confirmation is not run. AP repairs 68 and introduces six, but is not retrospectively relabeled as the primary candidate. The protocol required at least 0.5 points of gain over Four, 0.3 over each trained control, 0.5 points of bad-prefix reduction, regression at most 0.5%, and capability losses no worse than one point. The later three-pipeline studies use separate protocols.

Execution includes 4,096 training updates and 34,366 top-level calls. The complete-record companion retains all intervals, audit-state transitions, attribute heterogeneity, initial failures and reconstruction checks. This pilot establishes that deployable correction is possible, not that a semantic basis has a repair advantage.

J Repeated Repair and Matched Context Reading

These studies follow the initial pilot under separately fixed protocols. They first check whether optimization stabilizes internal-exchange training, then isolate the information read by the correction network. The later results do not change earlier stopping decisions.

J.1 Evidence identity and evaluation

All comparisons in this section freeze one task-adapted complete Qwen3-VL-30B-A3B-Instruct model and evaluate the same previously opened 128 images. The full domain has 2,923 inputs; selected-prefix completion has 1,665. Three trajectories vary correction initialization or paired training order as specified below, not the underlying VLM adaptation. The original material errors occupy only three inputs in two images, so broad two-attribute repair is not established by the total panel size. Pixel composites are auxiliary supervised training inputs, not members of the natural legal audit domain.

Accuracy is averaged over inputs within each image and then across images. A bad prefix contains at least one wrong answer; its rate is averaged across prefixes within each image, then across images. Regression is the fraction of originally correct inputs made incorrect, again averaged by image. These quantities need not track raw error counts. Reported mean counts average the three fixed trajectories. For paired intervals, trajectory differences are first averaged within image and images are resampled 10,000 times. The contextual-architecture and continuation studies use seeds 37920260924 and 38220260925, respectively. Intervals describe reused development data; they are not confirmation tests after selection.

J.2 Internal exchange training and optimization controls

At 512 updates per attribute, reducing Full-swap's learning rate from 10^{-3} to $3\,10^{-4}$ reduces mean regression from 0.852% to 0.214%. However, it does not establish an increment over ordinary paired CF, consistency or Pixel training. Table 19 retains the matched low-rate comparisons. Only the low-rate Full fits were newly run for that comparison; the other controls use locked historical outputs, not new repetitions.

Table 19: Low-rate correction comparisons on the reused development panel. All rows except Four average three paired trajectories. Percentages are image-equal; error counts are trajectory means.

Training	$\mathcal{D}$ accuracy	$\mathcal{U}^\star$ bad prefixes	Regression	Repair	New
Four	96.922	4.844	0	0	0
CF	98.741	2.956	0.175	51.67	5.33
Consistency	98.738	2.595	0.168	50.00	4.00
Full-swap	98.770	2.426	0.214	53.33	6.00
Pixel	98.850	2.478	0.162	56.33	5.33

A separate 64-step study compares a natural two-condition maximum loss with an internal four-condition maximum. Its 32 evaluation images are held out from that correction fit, but were included in initial VLM adaptation; they are not new external images. The evaluated union of 64 completed prefixes contains 977 inputs, not the full $\mathcal{D}$ or the older $\mathcal{U}^\star$. Natural/internal mean accuracy is 96.679/96.701%, versus 94.824% before correction; regression is 1.626/1.562%. Both learn, but unstable regression prevents expansion of the internal hard-maximum recipe. This stopping decision is not revised by later architecture experiments.

J.3 Contextual architecture and common objective

Each channel normalizes the unmodified 2,048-dimensional state by its token RMS, then applies a width-16 linear stem and GELU. Token uses a width-64 fusion layer on the current encoding. Generic concatenates that encoding with three learned softmax-attention summaries and uses a width-16 fusion layer. Structured replaces those learned summaries with means over the two 25-position task groups and the remaining 119 positions. Each fusion is followed by GELU and a rank-four output

Table 20: **Contextual correction helps, but task grouping adds no established benefit.** Reused development panel: 128 images, 2,923 full-domain inputs and 1,665 completion inputs. Rates are percentages; regression means previously correct inputs made wrong. Bad-prefix rate is defined in the text. The lower block has 512 extra updates and is a separate matched comparison, not an equal-cost competitor to the upper block.

Correction	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad prefixes	Regression	Repair	New
Four (before repair)	96.922	4.844	0	0	0
Token	98.596	2.891	0.264	52.67	7.33
Generic	98.881	2.348	0.251	54.67	7.00
Structured	98.056	3.108	0.360	37.00	11.00
Generic continuation	99.315	1.328	0.198	66.67	7.67
With task-group prior	99.316	1.328	0.194	66.33	7.67
With matched-group prior	99.299	1.458	0.170	66.33	6.67

layer. Parameter counts per attribute are 136,528, 135,760 and 135,568. All summaries are computed before simultaneously writing the 50 task positions. Output layers start at zero; stems are shared at initialization within each paired trajectory. The VLM, task adapter and write bases remain frozen.

The training pool contains 64 images and 965 distinct natural inputs. Each attribute/trajectory uses 256 pairs generated by the fixed ordering rule, with 1,536 Pixel composites across the six schedules. For natural receiver/suffix variants A, B and a Pixel composite P, the common objective is

$$\mathcal{L} = \tfrac{1}{2}(\mathrm{CE}_A + \mathrm{CE}_B) + 0.3\,\mathrm{CE}_P + 0.3\max(\mathrm{CE}_A, \mathrm{CE}_B) + 0.01\,L_{\mathrm{update}}. \tag{9}$$

CE is binary No/Yes cross-entropy. Ties in the maximum select the first term. The update penalty averages squared requested update norms over channels and written positions, then over receiver, suffix and donor states; Pixel adds no additional penalty term. AdamW uses batch one, zero weight decay, epsilon 10^{-8} and gradient clipping at one. Checkpoints preserve optimizer and random states. Training does not use internal four-condition or Full-swap losses here.

J.4 Selection and recipe-fixed repeats

Trajectory 17 screens learning rates $3\,10^{-4}$ and 10^{-3} at 128, 256 and 512 updates, giving 18 two-attribute candidate systems. A candidate must have overall accuracy loss at most 0.2 percentage points relative to Four, regression at most 0.5%, and per-attribute, root and target-change-pair accuracy losses at most one point. Within each step count and architecture, eligible learning rates are ordered by lower bad-prefix rate, higher accuracy, lower regression and finally lower learning rate. Among common step counts with an eligible candidate for all arms, selection uses the mean bad-prefix rate, then mean accuracy, mean regression and fewer steps. This selects 512 updates for all arms: $3\,10^{-4}$ for Token and Structured, and 10^{-3} for Generic. Trajectories 29 and 43 use those recipes without re-selection. Table 21 reports these two runs separately.

Table 21: Recipe-fixed trajectories 29/43 only, separate from the trajectory used for selection. Same reused development images; not an independently selected confirmation panel.

Correction	$\mathcal{D}$ accuracy	$\mathcal{U}^\star$ bad prefixes	Regression	Repair	New
Token	98.606	2.695	0.311	53.50	9.00
Generic	98.911	2.077	0.254	55.00	8.00
Structured	97.989	3.216	0.485	36.50	14.50

Structured does not meet the expansion rule of safety plus improvement over both controls with consistent repeat directions. All three architectures nevertheless improve on Four. Structured's accuracy is 0.825 percentage points below Generic (descriptive paired-image interval [-1.470,-0.293]); the two recipe-fixed repeats retain that direction. The protocol completes 12,288 formal updates, 188,460 top-level calls and 12.007 aggregate GPU hours, including engineering and retries. Calls and update counts are not exact FLOPs.

Table 22: **The final repair bases also undergo the exchange test.** Frozen bases from pipelines 17/29/43; no refitting or selection. Change/retain/unseen are semantic-target agreement (%) on 24 images disjoint from construction and repair training. The first three columns show ranges across pipelines. Last column lists hard-answer differences from correct/error donors in the same 28 third receivers. Full-slot's historical value is 7/28. Full per-pipeline and attribute results are in Appendix K.

Final space	Change	Retain	Unseen pairs	Donor differences
AP	0.00–1.39	100.00	33.33–34.72	0/0/1 of 28
Semantic	93.06–95.49	96.53	94.44–97.22	1/1/2 of 28
Predictive	0.00–1.74	100.00	33.33–35.07	1/0/0 of 28
Joint	0.35–0.69	99.65–100.00	33.33–34.03	0/0/1 of 28

J.5 A REMOVABLE PRIOR ON A STRONG GENERIC CORRECTION

Each corresponding Generic checkpoint receives 512 additional updates at $3\,10^{-4}$ with a fresh optimizer and common objective/order. The only added variable is a learned scalar attention-logit offset for each channel/summary, initialized to zero. Task uses groups of sizes 25/25/119; Matched uses fixed hash-selected arbitrary groups of identical sizes. Zero does not apply offsets. All three start with identical outputs and retain unrestricted attention. This tests a removable preference rather than enforcing group means. There is no endpoint or learning-rate re-selection. The results in Table 20 do not establish an increment from the task prior. The 18 fits complete 9,216 updates, 80,430 calls and 5.618 aggregate GPU hours. Verification reconstructs statistics from saved input records but reuses some analysis helpers, so it is not an independent software implementation.

The final reserved 256-image panel is not used in any result reported here. Appendix L reports the subsequently completed, separately fixed write-space comparison, including its new development panel.

K VALIDATING THE ACTUAL BASES USED FOR REPAIR

The early exchange map and the bases in the final repair comparison are different fitted objects. We therefore validate every final AP, Semantic, Predictive and Joint basis in pipelines 17, 29 and 43. All checkpoints and test examples are frozen before this additional inference. No basis, repair checkpoint, amplitude or primary criterion is changed in response.

The exchange test reuses the 576 fixed conditions on 24 images from Appendix G, with 288 conditions per attribute. These images do not overlap the 48 basis-construction images or the 64 repair-training images. The construction examples use pair types 00/11; the exchange test also includes 01/10. Joint uses no exchange construction, so "unseen" denotes this shared pair-type test stratum, not a separately matched exchange-training split. The panel was already available during earlier research and is not a new independent-image confirmation.

For fixed-space arms, we load the basis directly from each step-512 repair checkpoint and verify exact identity with its construction output. For Joint, we apply the deployment QR and sign convention to the final trainable raw basis on the GPU; the stored initial reader basis is not used in its place. All four orthonormal 2048×4 bases act on the same task positions. Exchange uses Equation 5 at unit strength, without invoking the learned correction reader. The subsequent repair results use the independently frozen deployment amplitudes. This tests the semantic interpretation of the write space, not whether the reader's natural update equals a donor replacement.

Table 22 summarizes the ranges shown in Figure 3c. Across pipelines, Semantic's image-equal target-change agreement is 93.87% [90.39,96.88], retention 96.53% [93.98,98.61], and unseen-pair agreement 95.60% [92.71,98.03]. Intervals resample 24 images 10,000 times with seed 39020260926, averaging pipelines within image; they describe this existing panel. AP, Predictive and Joint have target-change agreement at most 1.74% in every pipeline. Table 23 pairs these scores with the already frozen new-image repair outcomes. Joint's better repair is therefore not evidence of better performance on this semantic exchange test; conversely, Semantic's strong exchange does not

Table 23: One-to-one correspondence between final-basis validation and repair. Exchange uses unit-strength donor replacement; repair uses its frozen confirmation amplitude (Semantic 0.5, others 1). Change/retain/unseen and accuracy/regression are percentages. Restore/damage count the 37 natural-error directions. Fixed counts different answers in 28 fixed third receivers. These are different populations, not per-image causal mediation estimates.

Space	Run	Change	Retain	Unseen	Restore/damage	Fixed	$\mathcal{D}$ acc.	Regr.
AP	17	1.39	100.00	34.72	4/1	0/28	97.855	0.620
AP	29	0.00	100.00	33.33	2/2	0/28	98.080	0.432
AP	43	0.00	100.00	33.33	5/0	1/28	98.391	0.266
Semantic	17	95.49	96.53	97.22	17/9	1/28	97.881	0.292
Semantic	29	93.06	96.53	94.44	18/8	1/28	97.783	0.203
Semantic	43	93.06	96.53	95.14	20/9	2/28	97.973	0.322
Predictive	17	0.00	100.00	33.33	6/4	1/28	97.690	0.287
Predictive	29	0.00	100.00	33.33	7/3	0/28	97.772	0.191
Predictive	43	1.74	100.00	35.07	6/2	0/28	98.006	0.454
Joint	17	0.69	99.65	34.03	8/2	0/28	98.092	0.457
Joint	29	0.69	100.00	34.03	9/2	0/28	98.319	0.102
Joint	43	0.35	99.65	33.33	4/3	1/28	98.611	0.524

Table 24: Final-basis exchange results by attribute. Each entry gives the image-equal percentage for pipelines 17/29/43; each attribute has 12 held-out images and 288 conditions per pipeline.

Space	Attribute	Change	Retain	Unseen
AP	material	0.00/0.00/0.00	100.00/100.00/100.00	33.33/33.33/33.33
AP	size	2.78/0.00/0.00	100.00/100.00/100.00	36.11/33.33/33.33
Semantic	material	95.14/88.19/93.06	98.61/97.92/96.53	96.53/90.97/95.14
Semantic	size	95.83/97.92/93.06	94.44/95.14/96.53	97.92/97.92/95.14
Predictive	material	0.00/0.00/2.08	100.00/100.00/100.00	33.33/33.33/35.42
Predictive	size	0.00/0.00/1.39	100.00/100.00/100.00	33.33/33.33/34.72
Joint	material	0.00/0.00/0.00	100.00/100.00/100.00	33.33/33.33/33.33
Joint	size	1.39/1.39/0.69	99.31/100.00/99.31	34.72/34.72/33.33

establish incremental repair benefit. Unequal parameter and construction budgets remain limitations of that comparison, rather than being removed by this validation.

We also rerun all 37 fixed natural-error cases, including both transfer directions and the 28 eligible third receivers, for every final space. These images are disjoint from both fitting pools. There are 130 conditions per space and pipeline. For Semantic, fixed-receiver hard differences occur in 1/28, 1/28 and 2/28 cases; nonzero original-target margin differences occur in 24/28, 22/28 and 25/28. Mean absolute differences are 0.772, 0.714 and 1.031 logits. Unlike the earlier map's 0/28 result, the final bases exhibit some hard transfer, but this remains substantially less common than their attribute-composition success. The two tests ask different causal questions.

The run retains 8,472 intervention records, all with valid No/Yes generations and matching scoring/generation first tokens. 273 original states are captured and replayed per pipeline. Self-exchange preserves each checked answer and margin within 0.01; inputs are not modified in place and unselected positions remain unchanged. All VLM parameters remain frozen, with zero parameter gradients. The experiment uses 18,630 top-level model calls and 0.587 aggregate GPU hours on three H100 devices, including model loading and controls. No training update, checkpoint selection or final reserved-panel access occurs. The reconstruction package retains per-condition results, checkpoint hashes, membership checks, costs and deterministic aggregation code.

L Matched Write-space Construction and Repair

We tested whether constructing a write space from internal intervention responses adds repair value beyond an untrained space or one learned from supervised prediction. The task-adapted Qwen3-VL-30B-A3B-Instruct and its existing adapter remained frozen. All arms used the same Generic

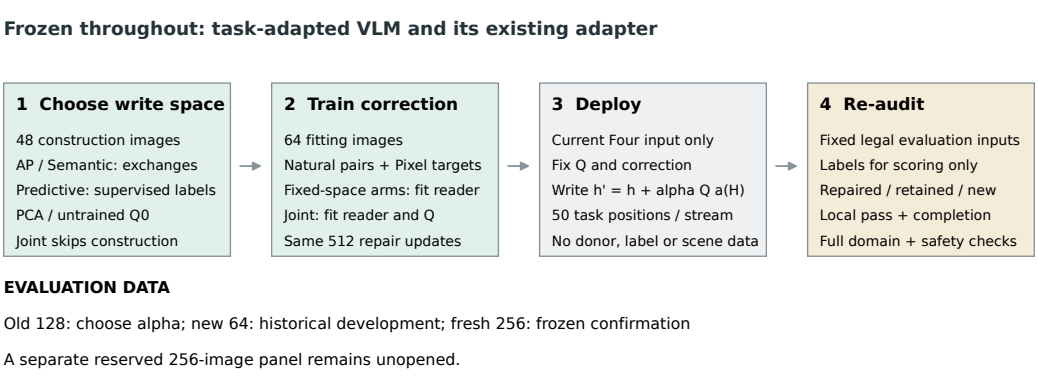


Figure 6: **From intervention evidence to donor-free repair.** Fixed-space arms construct Q before fitting the common reader a_ϕ; Joint learns both together. The VLM and task adapter remain frozen. Deployment uses only the current Four input. Historical comparisons use amplitude one; confirmation fixes the amplitude selected on development data before querying new images.

Table 25: **Repair improves accuracy but can leave support violations.** Same Generic correction and 512 updates; 64 development images. Rates (%) are image-equal means over three paired pipelines. Repair/New are mean input counts. AP is the historical primary; Semantic is secondary. Joint is a follow-up direct-learning control, not capacity- or total-compute-matched. Last column: completion misses/local passes for runs 17, 29, 43. Historical arms and repeat-level counts remain in Tables 26–27; Joint details are in Table 31.

Space	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regr.	Repair	New	$\mathcal{U}^\star$ miss/pass
Four	96.753	5.130	0.000	0.00	0.00	3/59 (fixed)
AP	97.954	2.813	0.199	25.00	3.67	2/60, 2/62, 1/61
Semantic	98.260	2.075	0.063	28.67	1.33	1/61, 1/62, 1/61
Joint	98.273	2.613	0.119	30.00	2.67	2/61, 2/61, 0/61

correction architecture, four rank-four write spaces, 50 task token positions and 135,760 trainable correction parameters per attribute. The primary candidate was answer-preserve (AP); the joint primary controls were its exact untrained initialization Q0 and Predictive. Semantic and PCA were secondary comparisons, not replacement primary candidates.

AP and Semantic used the same 48-image construction pool, 187 unique inputs and 288 examples per attribute. AP optimized the original receiver answer after internal exchange; Semantic optimized the attribute-composed target. Each received 576 updates. Predictive used the unmodified receiver and donor states to predict the seven binary labels defined below; its temporary readout was discarded before repair. PCA used centered per-channel covariance of the same task-token states, without output-based selection. Within a pipeline, AP, Semantic and Predictive began from the same Q0.

L.1 What the Predictive Control Learns

Let b_0^r, b_1^r and b_0^d, b_1^d be the two task attributes of receiver r and donor d. Metal and large are encoded as one; rubber and small are zero. Let $m \in \{(1,0),(0,1),(1,1)\}$ specify the exchanged task slots, and define $\widetilde{b}_i = m_i b_i^d + (1-m_i) b_i^r$. Predictive has seven binary targets, in this order:

$$\big(\hat{f}(r),\ \hat{f}(d),\ b_0^r,\ b_1^r,\ b_0^d,\ b_1^d,\ \mathbf{1}\{\widetilde{b}_0 = \widetilde{b}_1\}\big). \tag{10}$$

The first two are stored original VLM predictions; the next four use training-scene attributes. The seventh is the answer implied by the swap. The common construction pool contains originally correct pairs, so the stored original predictions equal the corresponding task targets there.

For each of the four feature streams, Predictive averages the 25 interior states of each task slot in each run. It projects these averages onto the current rank-four basis and divides by a fixed stream-specific RMS computed from unique construction-training task states. Concatenating receiver and

donor codes yields $2 \times 2 \times 4 \times 4 = 64$ values; the two-bit mask gives 66 inputs. A $66 \to 32 \to 14$ multilayer perceptron with GELU activation returns seven pairs of logits. The objective is their unweighted mean binary cross-entropy. The temporary head has 2,606 parameters; only the four bases, with 32,768 learned scalars before orthonormalization, are retained for correction. Predictive receives the same pairs and order and 576 updates at learning rate 0.02. Unlike AP and Semantic, it does not backpropagate through patched VLM executions. It controls for supervised construction, not for every possible use of extra VLM backpropagation.

L.2 Pixel Composites and Correction Training

Each natural Four input is a 408×408 canvas containing 192×192 tiles. Task slots 0 and 1 start at $(8 + 200i, 8)$ for $i \in \{0, 1\}$. For the prescribed swap mask, Pixel copies each selected task tile from donor to receiver at the same position, including its tile label. Every pixel outside the copied rectangles is unchanged. The other task tile, if not selected, and both suffix tiles remain from the receiver. The target is $\mathbf{1}\{\widetilde{b}_0 = \widetilde{b}_1\}$, not necessarily either original answer.

Donors query the same attribute but come from a different repair-training image. Among eligible receiver examples, a fixed SHA-256 ordering selects the donor while preserving the prescribed target-change stratum. The three schedules use seeds 17, 29 and 43; each has 256 pairs per attribute. Thus 1,536 composites are shared by all arms. These cross-scene montages need not represent a legal route in any single scene. They supply auxiliary supervision only and never enter the natural-domain audit denominators.

Each frozen space then received a zero-output-initialized Generic correction trained for 512 updates at learning rate 10^{-3}. The five arms shared all correction initial parameters, examples and their order within a pipeline. The objective was $\frac{1}{2}(\mathrm{CE}_A + \mathrm{CE}_B) + 0.3\mathrm{CE}_P + 0.3\max(\mathrm{CE}_A, \mathrm{CE}_B) + 0.01L_{\mathrm{update}}$. Training used the existing 64-image repair set, 965 natural inputs and 1,536 auxiliary Pixel inputs across the three schedules. Pixel examples were not included in the natural legal audit domains. Pipelines 17, 29 and 43 repeated basis construction and correction training; they were not independently trained VLMs. All 30 correction fits were evaluated at the fixed final endpoint. There was no learning-rate or checkpoint selection in this comparison.

L.3 Data Roles and Evaluation

The old development panel contained 128 images, 2,923 full-domain inputs and 1,665 completion inputs. A separately selected, output-blind new development panel contained 64 images, 1,569 full-domain inputs and 936 completion inputs. Neither panel was final confirmation. We report them separately rather than pooling their results. The new panel was balanced across attribute and root target, with 16 images in each cell. Under an explicit data-access amendment, an isolated preparation process read the combined scene file, removed registered used/reserved members and exported eligible scenes before output-blind selection. Final reserved images and predictions were not opened; we do not claim that their scene bytes were never read.

For each image, let E and C count Four's originally wrong and correct inputs, and let R and N count repaired and newly wrong inputs. Conditional repair is the mean of R/E over images with $E > 0$; regression is the mean of N/C over images with $C > 0$. These conditional populations differ. Accuracy is image-equal; bad-prefix rates additionally weight prefixes equally within an image. Invalid generations count as errors. Images without a target-changing legal neighbor pair are not evaluable for that metric, not assigned zero. There were 106 evaluable old images and 51 evaluable new images. Tables 26 and 27 provide both panels and every paired trajectory.

Table 28 retains the paired intervals for the historical primary and secondary contrasts.

The old Four baseline made 82 errors across 21 images; the new baseline made 64 errors across 12 images. On the new panel, material errors comprised 17 inputs from only two images, whereas size errors comprised 47 inputs from ten images. The overall and size support requirements were met, but the predefined five-image requirement for a material-specific repair claim was not.

Before evaluation, AP was required to satisfy the same improvement route against both Q0 and Predictive. The more-repair route required a completion bad-prefix reduction of at least 0.5 percentage points and a conditional repair increase of at least five points, while allowing at most 0.1-point in-

Table 26: Matched write-space comparison. Rates are percentages averaged over images and then three paired construction–repair pipelines. B is the image-equal bad-prefix rate in selected-prefix completion $\mathcal{U}^\star$; G is regression conditional on originally correct inputs; R_{cond} is the image-equal repair fraction conditional on an image containing a Four error. Repair/New are mean input counts per pipeline. AP was the primary candidate; Semantic and PCA were secondary. Old and new panels are development data, reported separately.

Panel	Write space	$\mathcal{D}$ acc.	B	G	R_{cond}	Repair	New
Old (128)	Four	96.922	4.844	0.000	–	0.00	0.00
	AP	98.249	3.086	0.356	53.663	42.00	10.67
	Semantic	98.640	2.587	0.325	64.793	51.00	10.33
	Predictive	97.992	3.281	0.241	43.656	30.67	6.33
	PCA	97.335	4.236	0.273	21.834	16.00	10.00
	Q0	97.381	4.214	0.113	21.517	13.33	3.67
New (64)	Four	96.753	5.130	0.000	–	0.00	0.00
	AP	97.954	2.813	0.199	61.169	25.00	3.67
	Semantic	98.260	2.075	0.063	65.451	28.67	1.33
	Predictive	97.812	3.915	0.035	47.801	21.33	0.67
	PCA	97.349	4.505	0.117	28.993	14.33	1.33
	Q0	97.106	4.679	0.011	21.007	8.67	0.33

Table 27: All fixed-endpoint pipelines. Each cell lists pipelines 17/29/43, not independent VLMs. No endpoint was selected using these outputs. Definitions match Table 26; rates are percentages.

Panel	Space	$\mathcal{D}$ accuracy	$\mathcal{U}^\star$ bad prefixes	Regression	Repair/New
Old	AP	98.39/97.69/98.67	2.89/4.13/2.24	0.20/0.60/0.26	42/5; 30/16; 54/11
Old	Semantic	98.88/98.42/98.62	2.30/2.83/2.63	0.22/0.38/0.38	55/7; 44/12; 54/12
Old	Predictive	98.08/97.62/98.27	2.89/4.26/2.70	0.08/0.29/0.35	26/3; 24/7; 42/9
Old	PCA	97.60/97.08/97.33	4.00/4.52/4.19	0.18/0.32/0.32	19/7; 11/10; 18/13
Old	Q0	97.76/96.95/97.43	3.67/5.30/3.67	0.05/0.25/0.04	19/3; 7/7; 14/1
New	AP	97.61/98.10/98.15	3.39/2.45/2.60	0.26/0.12/0.21	19/6; 26/3; 30/2
New	Semantic	98.17/98.25/98.37	2.01/1.82/2.40	0.08/0.10/0.00	26/2; 29/2; 31/0
New	Predictive	97.73/97.66/98.05	4.09/4.35/3.31	0.00/0.10/0.00	22/0; 18/2; 24/0
New	PCA	97.32/97.33/97.39	4.35/4.35/4.82	0.18/0.00/0.18	15/2; 11/0; 17/2
New	Q0	97.45/96.88/96.99	4.09/5.13/4.82	0.00/0.03/0.00	16/0; 4/1; 6/0

creases in regression and losses in full-domain accuracy. The safer-repair route required at least a 0.1-point and 25% relative regression reduction, with at most 0.1-point accuracy/bad-prefix deterioration and five-point conditional-repair loss. Both routes also required strict benefit directions and the specified noninferiority bounds in pipelines 29 and 43, Four-relative capability protection, an overall conditional-repair floor of 50%, and continuity on the old panel. These were development risk rules, not significance tests.

AP reduced new-panel bad-prefix rates relative to Q0 and Predictive by 1.866 and 1.102 points and increased conditional repair by 40.162 and 13.368 points, respectively. Those benefits had the same direction in both recipe-fixed repeats. However, mean regression increased by 0.187 and 0.164 points, exceeding the 0.1-point allowance. Pipeline 43 alone increased regression by 0.208 points against both. Its material target-change pair accuracy also decreased by 1.02778 points, slightly exceeding the one-point capability limit. Even without this small capability violation, the relative-regression criterion would fail. AP satisfied the old-panel continuity checks, but neither prespecified route passed on the new panel.

Image-paired percentile bootstrap intervals used 10,000 draws and seed 38320260925, carrying the paired pipelines together. For AP minus Predictive on the new panel, the differences were $+0.142[-0.246, 0.559]$ points in accuracy, $-1.102[-3.116, 0.408]$ in bad-prefix rate, $+0.164[0.0003, 0.378]$ in regression and $+13.368[1.389, 26.389]$ in conditional repair. The corresponding AP-minus-Q0 differences were $+0.848[0.204, 1.634]$, $-1.866[-4.818, 0.347]$, $+0.187[0.032, 0.395]$ and $+40.162[26.502, 55.208]$. These are descriptive development intervals, not confirmatory or multiplicity-adjusted claims. Conditional-repair intervals concern the twelve new images with baseline errors.

Table 28: **More repair does not establish lower risk.** AP minus each primary control on the new panel, in percentage points. Brackets are descriptive 95% paired-image bootstrap intervals (10,000 draws; pipelines carried together). Conditional repair uses the 12 baseline-error images; other rows use all 64. No confirmatory or multiplicity-adjusted inference is claimed.

Metric (preferred direction)	AP − Predictive	AP − Q0
$\mathcal{D}$ accuracy (↑)	0.142 [-0.246,0.559]	0.848 [0.204,1.634]
$\mathcal{U}^\star$ bad prefixes (↓)	-1.102 [-3.116,0.408]	-1.866 [-4.818,0.347]
Regression (↓)	0.164 [0.0003,0.378]	0.187 [0.032,0.395]
Conditional repair (↑)	13.368 [1.389,26.389]	40.162 [26.502,55.208]

Semantic had better mean new-panel repair, regression and bad-prefix rates than AP. Its three pipelines removed 26/29/31 errors and introduced 2/2/0. Nevertheless, pipeline 43 changed one previously correct material root on the old panel, a $1/64 = 1.5625$-point attribute-level root loss, above the one-point limit. It therefore cannot be promoted to a jointly safe winner, or retrospectively called the primary candidate. Repair also did not remove the need for re-auditing: AP left full-domain errors in 4/60, 6/62 and 4/61 locally passing new-image records; Semantic left 5/61, 4/62 and 3/61. Repeated records are not independent images.

Execution used 191,512 top-level calls and 12.688 aggregate GPU hours, within the authorized 285,000-call and 24-hour limits. This includes 6,912 differentiable AP/Semantic construction calls and 46,080 repair training calls. Predictive construction used 3,456 additional updates without patched VLM calls. Generation comprised 69,200 calls and 138,400 tokens; peak reserved memory was 60,446 MiB on each worker. Calls are not exact FLOPs; construction cost was not granted back to the controls. A second CPU computation reproduced all 2,880 model–image metric records, and checkpoint, order and weight-identity checks passed. This is computational verification, not an external blinded review. The final outcome is a measured repair–regression trade-off, not validated mechanism-guided safe repair.

L.4 What This Comparison Does and Does Not Isolate

The fixed recipe measures a construction effect, not a ranking after equal method-specific tuning. AP constrains responses to donor exchanges; deployment instead makes additive updates. Safety does not transfer automatically between these operations. Appendix N adds the current Generic reader with a jointly learned space. Appendix M evaluates Pair and fixed-canvas suffix blanking on the same panels. These controls test direct learning and simpler input changes; they do not match all construction compute or establish equal-risk superiority. An equal-backpropagation construction without exchange targets, equal-achieved-risk tuning remains untested. Appendix O subsequently applies a common-risk amplitude rule and confirms the frozen choices on new images; it does not retroactively change this development comparison.

M Same-panel Input-level Controls

These controls test whether the task can be repaired more simply by removing the two irrelevant tiles. They use the same frozen task-adapted VLM and the exact old/new development memberships of Appendix L, without additional fitting. Both interfaces are specified before their outputs are opened.

Pair crops the original 408×408 Four montage to its upper 408×208 task row and uses the existing two-tile prompt. The VLM processor may resize this image differently. Blank instead retains the Four canvas and prompt, replacing the two lower 192×192 tiles with the canvas background, RGB $(245, 245, 245)$. Raw task pixels and processed task-interior patches are checked for exact equality with Four. This is an input-interface change, not the internal replacement used to construct or apply the learned correction.

For each scene, attribute and ordered task pair, transformed pixels and text are identical across all suffixes. We verify that identity before querying each unique input once, then map its prediction back to the original legal routes. There are 1,243 distinct task pairs across both panels. Each interface uses 1,243 full scoring calls and 1,243 actual generations, for 4,972 calls across both controls. All

4,492 route records remain in the evaluation denominators. No invalid generations occur; the rule would count them as errors.

Table 29: **Removing context enforces invariance, not correct answers.** Same old/new development panels as the matched write-space study. Rates are image-equal percentages; B additionally weights completion prefixes equally within an image. Repair/New are raw input counts. These are two frozen input interfaces, not three training repetitions.

Panel	Interface	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ B	Regression	Repair	New	$\mathcal{U}^\star$ miss/pass
Old128	Four	96.922	4.844	0	0	0	2/117
	Pair	93.804	6.064	4.208	16	110	0/114
	Blank	95.415	4.390	2.320	12	57	0/116
New64	Four	96.753	5.130	0	0	0	3/59
	Pair	96.837	2.656	1.473	26	35	0/58
	Blank	96.324	3.880	2.173	30	45	0/57

Table 29 shows substantial regression in both interfaces. The new-panel accuracy differences from Four are $+0.085[-1.349, 1.511]$ percentage points for Pair and $-0.429[-2.327, 1.321]$ for Blank. On the old panel, they are $-3.118[-5.299, -1.230]$ and $-1.507[-2.773, -0.373]$. Intervals use 10,000 paired-image bootstrap draws with seed 38820260926 and are descriptive development intervals. Image-equal accuracy and raw input counts weight scenes differently; Pair's small positive new-panel accuracy difference is therefore compatible with more introduced than repaired inputs. Both interfaces exceed the 0.5% regression allowance on both panels.

Neither interface has a selected-prefix miss after local passage: suffix variation has been removed by construction, and each tested prefix retains a correct local anchor. This zero is an interface property, not a successful repair of Four's internal dependence. Full-domain errors remain after local passage: Pair has 13/114 old and 7/58 new records; Blank has 13/116 and 5/57. They occur on task pairs not validated by the local pass. The comparison separates enforcing support consistency from recovering correct predictions and does not identify the cause of the regressions.

N Directly Learning the Write Space

The fixed-space comparison leaves a practical alternative: learn the directions together with the correction, without constructing them through internal exchanges. We test this alternative with the current Generic reader, rather than treating the older linear Free correction as its substitute. This additional control is fixed after the earlier five-arm study; it does not change that study's primary comparison.

Joint starts from each corresponding Q0 and the identical Generic initialization. Each stream has a trainable matrix $W_c \in \mathbb{R}^{2048\times4}$; positive-diagonal QR supplies the orthonormal Q_c in Equation 6 at every forward pass. The reader and W_c are optimized together; the VLM, task adapter and processor stay frozen. Only the same 50 task-interior positions are written, using summaries computed before any state is changed. Joint has 168,528 trainable parameters per attribute, compared with 135,760 for a fixed-space correction. It is a direct-learning control, not a parameter-matched or total-compute-matched ablation.

The two attributes and seeds 17/29/43 give six fits. They reuse the exact natural inputs, Pixel composites, objective, optimizer, sample order, learning rate 10^{-3} and 512-update endpoint of Appendix L. No endpoint or learning rate is selected. Both development panels retain all 4,492 inputs. We report each run, all three together and the 29/43 subset; the latter is not an independent-image confirmation. Bootstrap differences use 10,000 paired-image draws and seed 38820260926, retaining repeated runs within images. These intervals are descriptive and not adjusted for multiple comparisons.

Eight-step engineering tests check zero-initialized identity, gradient startup, unchanged non-task states, absence of in-place input edits and frozen-model gradients. Restoring step four reproduces step eight exactly, including optimizer state. Cached replacement and live hooks agree on predictions, with target margin difference at most 0.01. Each evaluated route receives both full scoring and

real generation. End-of-run CPU checks compare all actual training inputs, not only the planned schedules, with the corresponding historical Q0 run. They also check optimizer and RNG state, basis orthogonality and scored checkpoint identity.

Table 30 reports the comparison; Table 31 retains every endpoint, and Table 32 gives new-panel paired differences. Joint improves image-equal accuracy over Four by 1.625 [0.778,2.552] percentage points on the old panel and 1.521 [0.642,2.589] on the new panel. Its mean regression is 0.256% and 0.119%, respectively. It satisfies the stated Four-relative safety checks on both panels: mean regression at most 0.5%, each run at most 1%, and the specified overall and attribute-level accuracy, root and target-change bounds. Passing these development thresholds is not a population-level safety guarantee. Joint also exceeds the separately reported 50% conditional-repair floor on each panel.

Table 30: **Preconstructed versus jointly learned write directions.** Rates are image-equal percentages; bad-prefix rates also weight prefixes equally within images. Repair/New are mean raw counts over three paired runs. Joint adds 32,768 learned scalars per attribute. Historical fixed-space runs are reused, not rerun or reselected.

Panel	Space	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regression	Repair	New
Old128	AP	98.249	3.086	0.356	42.00	10.67
	Semantic	98.640	2.587	0.325	51.00	10.33
	Predictive	97.992	3.281	0.241	30.67	6.33
	Q0	97.381	4.214	0.113	13.33	3.67
	Joint	98.548	2.878	0.256	50.33	6.67
New64	AP	97.954	2.813	0.199	25.00	3.67
	Semantic	98.260	2.075	0.063	28.67	1.33
	Predictive	97.812	3.915	0.035	21.33	0.67
	Q0	97.106	4.679	0.011	8.67	0.33
	Joint	98.273	2.613	0.119	30.00	2.67

Table 31: **Every Joint endpoint and its re-audit.** No endpoint selection. Repair/New count inputs; the final column counts locally passing image records containing a completion error.

Panel/run	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regression	Repair	New	$\mathcal{U}^\star$ miss/pass
Old / 17	98.455	3.307	0.547	54	12	3/121
Old / 29	98.202	3.281	0.062	33	2	2/120
Old / 43	98.986	2.044	0.158	64	6	1/123
New / 17	98.102	2.526	0.313	28	7	2/61
New / 29	98.240	3.229	0.046	29	1	2/61
New / 43	98.478	2.083	0.000	33	0	0/61

On new images, AP minus Joint is $-0.319[-0.651, -0.040]$ points in accuracy. AP repairs fewer inputs than Joint in all three runs, although its bad-prefix and regression differences change direction across runs. Semantic has nearly the same mean accuracy as Joint, with fewer bad prefixes and less regression, but these descriptive intervals include zero (Table 32). The 29/43 subset also does not establish a consistent Semantic advantage: its accuracy differences are $+0.005$ and -0.111 points, and its bad-prefix differences are -1.406 and $+0.313$ points. An interval containing zero does not establish equivalence.

Joint's new-panel repairs concern 9/11/12 distinct images, and its new errors concern 1/1/0 images in runs 17/29/43. Some images contain both types of transition. Its remaining completion misses are 2/61, 2/61 and 0/61 local passes. The zero in run 43 is not selected as the reported model, and its full-domain re-audit still finds errors in 2/61 local passes on untested task pairs. The other runs have 3/61 and 4/61 full-domain residuals. All runs and transitions are retained.

These results support direct repair without preconstructed intervention directions under the tested recipe. They do not rule out advantages at another risk level, with equal total construction cost, or after equal method-specific tuning. Joint has more learned correction parameters, whereas AP and Semantic incur separate construction costs. Neither difference is erased by matching repair updates.

Table 32: **Paired-image differences from Joint on new development images.** Units are percentage points. Descriptive 95% intervals use 10,000 image bootstrap draws with trajectories kept together. Conditional repair concerns twelve images with original errors.

Comparison	Metric	Difference	95% interval
AP − Joint	$\mathcal{D}$ accuracy	-0.319	[-0.651, -0.040]
	$\mathcal{U}^\star$ bad prefixes	0.200	[-0.660, 1.085]
	Regression	0.079	[-0.074, 0.273]
	Conditional repair	-5.150	[-14.352, 4.340]
Semantic − Joint	$\mathcal{D}$ accuracy	-0.013	[-0.333, 0.340]
	$\mathcal{U}^\star$ bad prefixes	-0.538	[-1.337, 0.130]
	Regression	-0.056	[-0.304, 0.111]
	Conditional repair	-0.868	[-6.424, 5.208]
Predictive − Joint	$\mathcal{D}$ accuracy	-0.461	[-1.047, 0.056]
	$\mathcal{U}^\star$ bad prefixes	1.302	[0.087, 2.908]
	Regression	-0.084	[-0.319, 0.070]
	Conditional repair	-18.519	[-29.630, -6.944]

The original primary advancement rule remains failed; this follow-up neither promotes Joint to a retrospective primary candidate nor constitutes independent confirmation.

Together with the Pair/Blank controls, this extension uses 41,348 top-level calls and 2.194 aggregate GPU hours, including engineering and interrupted attempts. There are 3,072 formal updates and backwards, 9,216 checkpointed training branches, 16,008 generation calls and 32,016 generated tokens. Peak reserved memory is 60,316 MiB. Calls do not count nested checkpoint recomputation as separate VLM invocations and are not exact FLOPs. A recorded resource-inheritance bug would have applied older runtime limits; an output-blind external budget guard was added before either declared limit was reached. Final ledger checks confirm completion within the declared 8-hour, 60,000-call limits. This engineering issue changes neither examples nor optimization.

O Risk-constrained Calibration and Frozen Confirmation

This follow-up tests a decision missing from the fixed-amplitude study: whether intervention-trained and directly learned corrections differ after applying the same development risk constraints. It trains no new weights, searches no layers and does not revise the earlier primary failure. The existing AP, Semantic, Predictive and Joint checkpoints at step 512 are frozen in three paired pipelines.

O.1 Calibration before confirmation outputs

For each method, deployment uses $h' = h + \alpha Q a_\phi(H)$ with $\alpha \in \{0, \frac{1}{2}, 1\}$. One amplitude is shared across both attributes and all three pipelines. Zero reuses Four; one reuses hash-checked historical outputs. Only one-half requires new inference on the old 128-image development panel. The scale multiplies the float32 update before addition and BF16 writeback, not the already rounded state difference.

A candidate must keep overall image-equal accuracy within 0.2 percentage points of Four, mean regression at most 0.5%, and each pipeline's regression at most 1%. Within each pipeline and attribute, accuracy, root correctness and target-changing both-correct rate must not drop by more than one point. Among eligible candidates, we minimize mean completion bad-prefix rate, then maximize accuracy, minimize regression and choose the smaller amplitude. Zero provides a defined no-update fallback; it is not counted as a successful repair.

The resulting amplitudes are AP 1, Semantic one-half, Predictive 1 and Joint 1. Semantic at one fails the material-root check in pipeline 43; the other amplitudes and comparisons are retained in the source record. An independent implementation reproduces all eligibility decisions and selections without reading confirmation predictions. This is equal-candidate, common-risk calibration, not exact matching of achieved risks, trainable capacity, construction compute or extensive hyperparameter tuning.

Table 33: All calibration candidates on the old 128-image panel. Rates (%) are image-equal means over three fixed pipelines. Eligibility includes all overall and per-attribute capability checks, not only the columns shown. The same amplitude applies to both attributes and all pipelines; no confirmation output enters this table.

Method	α	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regr.	Repair	Eligible	Selected
AP	0	96.922	4.844	0.000	0.00	Yes	No
AP	0.5	97.621	4.063	0.221	21.67	Yes	No
AP	1	98.249	3.086	0.356	42.00	Yes	Yes
Semantic	0	96.922	4.844	0.000	0.00	Yes	No
Semantic	0.5	97.916	3.455	0.119	28.00	Yes	Yes
Semantic	1	98.640	2.587	0.325	51.00	No	No
Predictive	0	96.922	4.844	0.000	0.00	Yes	No
Predictive	0.5	97.455	4.128	0.114	17.33	Yes	No
Predictive	1	97.992	3.281	0.241	30.67	Yes	Yes
Joint	0	96.922	4.844	0.000	0.00	Yes	No
Joint	0.5	97.908	3.281	0.064	26.67	Yes	No
Joint	1	98.548	2.878	0.256	50.33	Yes	Yes

Table 33 retains every calibration candidate, including the excluded Semantic endpoint and the no-update fallback.

O.2 New images, fixed model and retained denominators

We select 256 images absent from the fitting and historical evaluation memberships, using only scene geometry and a fixed hash order. Material/size and root target No/Yes form four cells of 64 images. Each scene has a 4–64-input legal domain; its chosen root has another completion and a legal one-position neighbor. No correctness or error count determines membership. The selected images provide 6,299 legal inputs, 1,327 local inputs and 3,607 selected-prefix completion inputs. The 2,280 missing completions cover 254/256 images. Thus the certificate's query burden can be measured before any output is observed.

These images come from an earlier filtered, nonreserved scene pool. That preparation had read combined scene metadata before excluding reserved members. We therefore do not claim that scene metadata had never been inspected; image pixels and model outputs had not been used for project selection. The separate final reserved256 images and outputs remain unopened. The present panel is a fixed-checkpoint, new-image confirmation, not independent pretraining, task adaptation or a second VLM.

All four methods are evaluated on all legal inputs after amplitude selection is locked. No image or checkpoint is dropped. Invalid generations count as errors; no target-changing edge makes that image's both-correct metric unevaluable rather than zero. Every evaluated input receives full scoring and actual greedy generation. Baseline calls are shared across methods, with cross-device replay checks on previously observed inputs.

O.3 Primary comparison and support requirements

The primary estimand is Semantic minus Joint in image-equal completion bad-prefix rate. Within each image, we first average the three paired pipeline differences. We then resample images 10,000 times with seed 38920260926 and lexicographic image order. The joint rule requires the 95% interval's upper endpoint for bad-prefix difference to be below zero, the regression upper endpoint to be at most 0.1 percentage points, and the accuracy lower endpoint to be at least -0.2 points. Four-relative safety checks also apply. At least five baseline-error images per attribute are required for a cross-attribute repair claim. This operational minimum is not a power guarantee or population-level safety certification.

The joint rule requires all three endpoints, rather than allowing selection of a favorable metric. Other comparisons are secondary. Intervals describe uncertainty over this geometry-selected image population with the trained model and pipelines held fixed; they do not capture uncertainty from

task adaptation, basis fitting or development selection. Per-attribute and per-pipeline results are reported even when the primary rule fails. The rule is conjunctive: no one endpoint substitutes for another. Secondary intervals are not multiplicity-adjusted and do not license additional confirmatory superiority claims.

O.4 Completed confirmation and all comparisons

Table 6 gives the primary aggregate report; Table 34 retains every fixed endpoint. The baseline makes 191 errors on 47 of the 256 images, including ten material and 37 size images. All methods pass all predefined Four-relative safety checks, but the primary mechanism-increment criterion fails: Semantic does not establish lower bad-prefix rate, accuracy noninferiority or regression noninferiority relative to Joint. The respective limits are an upper endpoint below zero, a lower endpoint at least -0.2 points and an upper endpoint at most 0.1 points. The observed intervals in Table 35 fail these limits. No endpoint, amplitude or candidate is changed after confirmation.

Table 34: All frozen confirmation endpoints. Percentages are image-equal. Miss/pass fractions count images within each run; runs share the same 256 images and VLM. No endpoint is selected on this panel.

Method	Run	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regr.	Repair	New	$\mathcal{U}^\star$ miss/pass	$\mathcal{D}$ miss/pass
AP	17	97.855	3.223	0.620	85	26	4/240	16/240
AP	29	98.080	3.477	0.432	87	34	4/237	14/237
AP	43	98.391	2.669	0.266	115	10	3/243	8/243
Semantic	17	97.881	3.737	0.292	74	11	7/240	16/240
Semantic	29	97.783	3.965	0.203	64	20	5/237	17/237
Semantic	43	97.973	2.897	0.322	94	13	3/243	14/243
Predictive	17	97.690	4.030	0.287	68	12	6/238	15/238
Predictive	29	97.772	4.128	0.191	57	16	7/238	20/238
Predictive	43	98.006	3.678	0.454	93	25	6/238	16/238
Joint	17	98.092	2.617	0.457	100	37	4/242	16/242
Joint	29	98.319	2.552	0.102	96	6	2/241	13/241
Joint	43	98.611	2.051	0.524	133	31	1/243	9/243

Table 35: Frozen confirmation differences in percentage points. Brackets are 95% paired-image bootstrap intervals from 10,000 draws, with pipelines averaged within each image before resampling. Semantic minus Joint is the primary comparison; other contrasts are secondary and are not substitutes for that endpoint.

Contrast	$\mathcal{D}$ accuracy	$\mathcal{U}^\star$ bad prefixes	Regression
AP–Four	1.069 [0.508,1.699]	-2.235 [-3.409,-1.181]	0.440 [0.189,0.755]
Semantic–Four	0.839 [0.423,1.310]	-1.825 [-2.875,-0.911]	0.272 [0.065,0.546]
Predictive–Four	0.783 [0.369,1.240]	-1.413 [-2.339,-0.608]	0.311 [0.130,0.555]
Joint–Four	1.301 [0.699,1.974]	-2.951 [-4.427,-1.638]	0.361 [0.169,0.591]
Semantic–Joint	-0.462 [-0.902,-0.089]	1.126 [0.293,2.209]	-0.089 [-0.339,0.176]
Semantic–Predictive	0.057 [-0.204,0.295]	-0.412 [-0.781,-0.076]	-0.039 [-0.217,0.150]
Semantic–AP	-0.230 [-0.551,0.071]	0.410 [-0.028,0.922]	-0.168 [-0.415,0.016]

The direction of the bad-prefix and accuracy differences favors Joint in every pipeline (Figure 4). The regression interval crosses zero; this is not evidence of equivalent risks. The three pipelines share both images and a VLM, so their individual estimates are shown without treating them as independent model replications. Table 36 separates the two attributes. For material, Semantic removes 5.33 old errors and introduces 8.00 on average, despite its pooled net repair. These raw counts and image-equal rates answer different questions and are both retained.

Per-input records retain repaired, persistent and introduced errors. Prefix transitions separately record whether a previously correct prefix becomes bad and whether an originally bad prefix is cleared. Table 37 decomposes the ten original completion misses. For example, AP clears seven in every pipeline, but one or two others become local failures rather than successful completion checks. Joint clears seven, eight and eight, respectively; one, one and one new misses arise outside the

Table 36: Confirmation by attribute, averaging three fixed pipelines. Rates are image-equal percentages; each attribute contains 128 images. Errors are Four errors. Repair/New are mean input counts. The error-image column counts distinct images, not image–run records.

Attribute	Method	$\mathcal{D}$ acc.	$\mathcal{U}^\star$ bad	Regr.	Repair	New	Error images
Material	Four	98.085	2.539	0	0	0	10
Material	AP	98.510	2.170	0.167	9.00	5.00	10
Material	Semantic	98.147	2.452	0.284	5.33	8.00	10
Material	Predictive	98.310	2.235	0.086	5.33	3.33	10
Material	Joint	98.799	1.237	0.047	15.33	2.67	10
Size	Four	95.995	8.177	0	0	0	37
Size	AP	97.707	4.076	0.712	86.67	18.33	37
Size	Semantic	97.612	4.614	0.260	72.00	6.67	37
Size	Predictive	97.335	5.655	0.536	67.33	14.33	37
Size	Joint	97.883	3.576	0.675	94.33	22.00	37

original cohort. The full transition records retain the origins of these new misses. The executable

Table 37: Where the ten original completion misses go after correction. Counts are images within each fixed pipeline. Repaired means all completion inputs become correct; Local fail means the case is no longer a local pass, not a certified repair. New misses arise outside the original ten-case cohort. The last column reconciles with Table 34. This is descriptive decomposition, not a replacement primary endpoint.

Method	Run	Repaired	Local fail	Still missed	New misses	Total misses
AP	17	7	2	1	3	4
AP	29	7	1	2	2	4
AP	43	7	1	2	1	3
Semantic	17	4	1	5	2	7
Semantic	29	5	0	5	0	5
Semantic	43	7	1	2	1	3
Predictive	17	5	1	4	2	6
Predictive	29	6	0	4	3	7
Predictive	43	6	2	2	4	6
Joint	17	7	0	3	1	4
Joint	29	8	1	1	1	2
Joint	43	8	2	0	1	1

support certificate covers Four and twelve corrected systems, or 3,328 image–system records. For Four, 232/234 local passes leave completion queries untested; 224/234 pass completion and ten yield a contradiction to a correct local anchor. The exact certificate is consistent with these outcomes. This secondary deterministic reconstruction uses the same predictions, not additional model calls or another confirmation test.

O.5 Constraint-removal control

Using only the old128 and new64 development scenes, we hold the root, model, question and first two relations fixed and remove the third relation. We enumerate all newly permitted one-edit neighbors: 725 additional inputs, evaluated with Four alone. The strict local set is a subset of the relaxed one. The original completion domain remains the reference for defining a miss, so a changing reference set cannot explain the comparison.

The focal cohort consists of strict local passes with a strict completion error, selected without relaxed outputs. We count how many cease to pass under relaxation and distinguish errors at locally anchored task pairs from unanchored errors. The former are direct support contradictions; the latter alone are not. Newly admitted inputs do not enter original-domain accuracy or repair denominators. The control measures how this experimental constraint changes available comparisons, not whether it is mandatory in practice.

Table 38: Removing the final relation adds one-edit inputs while leaving the original completion reference fixed. The cohort is defined before relaxation by a strict local pass and a completion error. "Anchored" means a new wrong input has the same task pair as a correct strict-local input. These controls do not belong to the original legal domain.

Panel	New inputs	Strict cohort	Exposed	Anchored	Still passes
New	261	3	0	0	3
Old	464	2	1	1	1

The added inputs expose one of two old-panel focal cases and none of three new-panel cases (Table 38). The exposed input preserves the root task pair and is answered wrongly. It was illegal under the strict contract, so this result does not mean the original two-edit witness itself became a one-edit route. It establishes a newly available informative comparison; four other strict misses remain invisible to the relaxed local check.

O.6 Equal-cost interpretation of constraint relaxation

The analysis in Table 5 was specified after the original outputs were available and is descriptive, not a new confirmation trial. For image i, let n_i be the strict local nonroot count and m_i the relaxed nonroot count. We compare all n_i strict inputs with a uniformly sampled n_i-subset of the m_i relaxed candidates, keeping the root in both. All 192 development images are retained. No budget, image, relation or attribute is selected using the resulting comparison.

If the root is correct and a pool of m inputs contains k errors, the exact discovery probability at budget b is $1 - \binom{m-k}{b}/\binom{m}{b}$. An incorrect root is already a discovery. For support contradictions we count subsets with two different outputs at the same task pair, including the root. For a nonroot group with n_0, n_1 answers, the polynomial for noncontradictory selections has constant coefficient one and degree-j coefficient $\binom{n_0}{j} + \binom{n_1}{j}$ for $j > 0$. For the root group, only selections agreeing with the root are allowed. Multiplying these polynomials and extracting degree b counts noncontradictory subsets; division by $\binom{m}{b}$ gives their probability. Independent brute-force subset enumeration verifies 160 small cases. Each logical input query requires scoring plus generation, so equal query totals also match top-level calls.

Only one of the five focal images contains a newly admitted wrong input. Its exact same-budget discovery probability is $15/22 = 68.18\%$, versus zero under the strict candidate set. Across all images, however, relaxed uniform sampling finds fewer errors and contradictions at the same cost. The extra candidates include correct answers and can dilute the chance of selecting existing errors. This distinguishes the existence of a formerly unavailable witness from an overall efficiency gain. It neither establishes that the constraints are practically mandatory nor changes the legal reference used for the original repair comparisons.

O.7 Engineering and resource accounting

All weights, inputs, processor files, scripts and selected amplitudes are hashed. Zero-update identity, non-task-state preservation, no in-place input modification, exact cached-state replay, and cached versus live-hook prediction agreement are checked before formal inference. Target-margin drift must be at most 0.01. The old-amplitude endpoints are replayed on each authorized GPU. The runtime is Python 3.10.20, PyTorch 2.11.0+cu128, Transformers 5.12.1, PEFT 0.19.1 and NumPy 1.26.4.

A shared atomic ledger counts attempts before execution and deduplicates committed evaluations by input, checkpoint, amplitude and protocol identity. The ceiling is 400,000 top-level calls and 12 aggregate GPU hours, including loading, engineering and interrupted attempts. No training update or backward pass is permitted. The amplitude lock precedes the first confirmation call. Statistical reconstruction independently checks geometry, labels, per-image metrics and paired intervals; it is not a second GPU experiment. The executable support-certificate check is a secondary deterministic analysis, not another prespecified superiority endpoint. It verifies the local-anchor equivalence and retains explicit witnesses, unqueried counts and binary extension bounds for every evaluated system.

Completed execution used 235,568 top-level calls and 7.252 aggregate GPU hours across GPUs 0, 2, 3 and 7, with no failed calls, training updates or backward passes. Of these calls, 117,784 perform actual generation, totaling 235,568 generated tokens. Peak reserved memory is 60,166 MiB. The total includes old-panel calibration, Four capture, constraint controls, confirmation and engineering. Historical fitting and direction-construction costs are additional, not erased by freezing. Independent reconstruction checks 698 reported metric cells, all 256 geometries and labels, all safety decisions, calibration selection, disjoint membership and the 725 constraint-control inputs.

For cost interpretation, the shared Generic correction uses 24,072,064 multiply–add pairs per intervention: 22,151,168 for the token stem, 64,896 for attention summaries, 204,800 for fusion, 12,800 for coefficients and 1,638,400 for the low-rank write. Counting a multiply and an add separately gives about 48.14 million linear-algebra operations. This analytical count excludes activation, normalization, softmax, copying, Joint's QR and the entire VLM. Scoring plus actual generation invokes the correction twice. It is neither a measured latency nor an exact total-FLOP estimate.

P Interpretation, Attribution and Remaining Limits

Unlike an instance-specific sufficient reason (Cooper & Amgoud, 2023), our support hypothesis must fit every queried group. Unlike longer-context distractor tests (Sharma et al., 2024), Four holds input length fixed. Neither distinction makes constrained distinguishability a new general principle.

Patching depends on granularity and metrics (Zhang & Nanda, 2023; Heimersheim & Nanda, 2024); the different interventions here do not constitute a layer-by-layer causal proof. ReFT and MoReS establish the representation-tuning foundation (Wu et al., 2024; Bi et al., 2024). Our experiment tests intervention-trained directions within that foundation. Historical safety failures and the new confirmation's failed incremental-benefit test bound the tested recipes, not the possibility of mechanism-guided improvement.