RULE-CONDITIONED ATTRIBUTE COMPARISONS FOR VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual rules can preserve a collection of attribute values across panels even when objects within each panel disagree. With exact perception, the appearance consensus summary maps every heterogeneous attribute inventory to one inconsistency state and can erase distinctions between answers. We introduce a rule-conditioned comparison layer for fixed probabilistic reasoners. Under eligible Constant rules, occupancy-weighted histograms retain the values that should persist. Under scalar appearance rules, conditioning candidate beliefs on ordinary values changes both their inconsistency penalty and categorical weighting. Perception, inferred rules and other score terms remain matched. On 2,200 fresh high-count I-RAVEN questions in a nine-slot layout, inventory comparison raises PrAE accuracy from 94.50% to 99.68% and NVSA accuracy from 93.77% to 99.50% with shared perception and prespecified low-count guardrails met. Reusing this cohort with three published visual checkpoints retains average gains of 4.74 and 4.23 points. A separately frozen 3,300-question low-count study confirms a 0.45-point conditioning gain. Exact information-loss analysis and matched controls identify retained values as the main inventory benefit, while crossed scalar comparisons separate the effects of candidate shape and consistency mass.

1 INTRODUCTION

Suppose a row of panels contains a light circle and a dark circle, and the circles move one position to the right at each step. Color is constant across the row: each circle keeps its color. Yet the objects within a panel do not share a color. A candidate with a different pair of shades can obey the same movement and object count while violating the color rule. The reasoner must retain *which colors are present*, rather than only whether all colors agree.

This distinction matters in probabilistic visual reasoning. PrAE (Zhang et al., 2021) converts object predictions into distributions over scene attributes, selects relations from context, and scores candidates against the executed predictions. Its appearance summary assigns probability to each event "all occupied objects have value v," with the remaining probability assigned to an inconsistency state. These are valid events under the model's factorization. They are not an attribute inventory. With exact perception, every heterogeneous inventory maps to the same inconsistency state, even when its values would distinguish the correct answer from a distractor.

We adapt candidate comparison to the inferred appearance rule. Under a Constant rule in an object-preserving spatial branch, occupancy-weighted pooling retains the attribute inventory for comparison with the backend's context reference. For scalar rules, we separately test conditioning candidate laws on ordinary values. Both changes require no fitting and preserve other score terms. The contribution is a rule-conditioned replacement of the affected comparison inside existing reasoners, with perception and execution held fixed.

Sharper perception cannot recover information discarded by a summary. Four fresh studies test inventory comparison, reasoning implementation and scalar conditioning. Presence, deletion and matching controls distinguish the benefits of values, frequencies and bindings; crossed scalar comparisons separate candidate shape from consistency mass.

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.

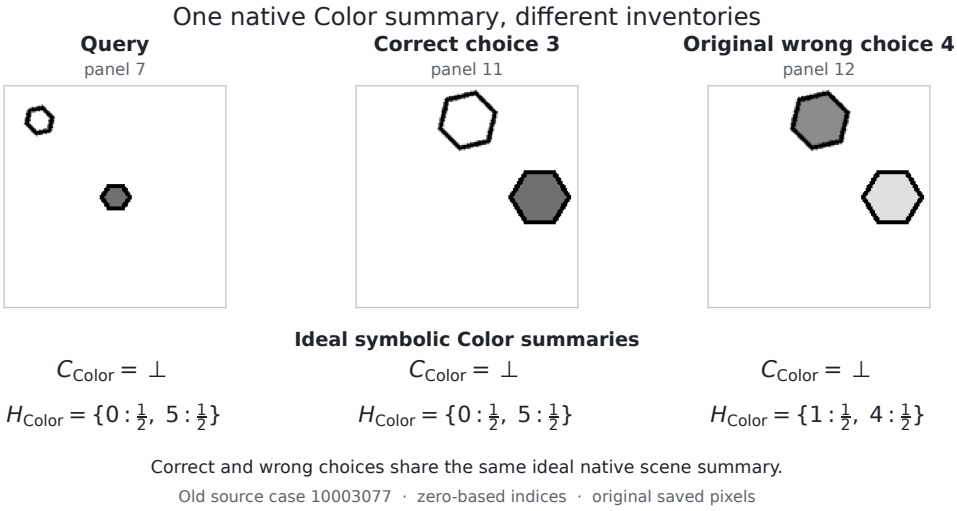

Figure 1: **Across-panel constancy does not require within-panel uniformity.** In this audited source question, query and correct candidate share two Color values; the distractor contains different values. Both map to ideal Color state $\perp$, with their other ideal native summaries also equal. Color inventories retain the distinction. Displayed equalities use saved symbolic annotations; measured soft scores need not coincide.

Our contributions are:

- A comparison adapter that preserves heterogeneous inventories under eligible Constant rules, with a separately tested conditioning extension for scalar rules.
- An exact characterization of consensus information loss, a conditional soft-prediction bound and a divergence decomposition separating ordinary-value shape from consistency mass.
- Four prospectively frozen studies with matched reasoning and fixed perception, plus transfer to three published visual checkpoints and complete censuses of corrections and harms.

2 Related Work

Object and relation representations. RAVEN provides structured visual matrices and annotations (Zhang et al., 2019); I-RAVEN changes candidate generation to reduce answer-set shortcuts (Hu et al., 2021). SCL factorizes learned reasoning over objects, attributes and relationships (Wu et al., 2020). OCRA and Slot Abstractors learn relational abstractions of object representations (Webb et al., 2023; Mondal et al., 2024). These approaches learn relational representations. Our matched comparison isolates which statistic of the same slot predictions is used to evaluate an already selected relation inside a probabilistic backend.

Probabilistic execution. PrAE combines probabilistic abduction with symbolic execution (Zhang et al., 2021); its published scene, planner and executor form our backend. NVSA also uses an inconsistency state for RAVEN appearance; its PGM extension already forms distributions over value combinations from clipped occupancy-weighted category sums (Hersche et al., 2023). Our H/P controls compare normalized per-value marginals, not that full combination distribution. The contribution is the relation-conditioned repair of RAVEN consensus comparisons and its matched evidence, rather than a first value-set representation. ARLC learns context-augmented vector-symbolic rules from metadata-provided attributes (Camposampiero et al., 2024). ALANS supplies the publicly released ObjectCNN weights for our visual-transfer test (Zhang et al., 2022). We test candidate comparison with perception, inferred rules and other score terms matched, rather than compare these full systems' reported benchmark accuracies.

Pooling and matching. Histograms, Jensen–Shannon divergence and optimal transport provide established ways to compare sets of objects. Sharing their inputs and inferred rules lets us test whether marginal inventories suffice or whether attribute pairings and spatial correspondence add predictive value. Section 3.3 separates the distinctions each comparison retains; the experiments measure their contribution to answers.

3 PRESERVING THE EVIDENCE REQUIRED BY A RULE

3.1 FIXED PERCEPTION AND NATIVE EXECUTION

A question has eight context panels and eight candidate answers in a fixed nine-slot layout. Panel indices start at zero: reference panel 7 is the last context panel (the query in Figure 1); panel $8 + y$ is the correct completion for answer index y. For panel i, slot s and appearance attribute $a \in \{\mathrm{Type}, \mathrm{Size}, \mathrm{Color}\}$, the perception interface supplies occupancy π_{is} and a proper categorical distribution q_{isa} on $\mathcal{V}_a$. The native scene engine forms Number, Position and appearance laws from these outputs. Context-based native abduction selects either a Position or a Number relation, together with an appearance relation for each attribute. Execution produces the predicted scene laws used to compare each candidate.

For appearance, the ideal native summary is

$$C_{ia}(v) = \sum_{M \neq \varnothing} P_i(M) \prod_{s \in M} q_{isa}(v), \qquad C_{ia}(\perp) = 1 - \sum_{v \in \mathcal{V}_a} C_{ia}(v), \tag{1}$$

where P_i is the modeled occupancy distribution conditioned on a nonempty panel. The regular coordinates describe the all-same events. The $\perp$ coordinate collects other configurations. Our mathematical analysis concerns this ideal law; the native implementation also uses probability floors and numerical clamps.

We retain the exact native float32 score as a reference. A separate normalized-JS baseline renormalizes each native law and computes natural-log Jensen–Shannon divergence in float64. Write its candidate score as

$$S_0(j) = -D_{\mathrm{NP}}(j) - \sum_a D_{Ca}(j). \tag{2}$$

D_{NP} is the selected Number or Position comparison, not their sum. D_{Ca} compares the native executed prediction with candidate j for attribute a. Within each inference variant, all comparison arms share these baseline terms. An eligible appearance term is the only replacement. Normalization of score inputs is distinguished from precision changes inside rule abduction, which are controlled separately below.

3.2 RULE-GATED ATTRIBUTE DISTRIBUTIONS

An attribute is eligible when its inferred appearance rule is Constant and the inferred Position rule preserves its inventory. The original gate includes Position Constant and Progression $\pm 1, \pm 2$; the expanded gate adds Distribute Three. Write the eligible attributes as $\mathcal{A} \subseteq \{\mathrm{Type}, \mathrm{Size}, \mathrm{Color}\}$. No annotations or candidate correctness determine this gate. These generator branches retain the object list and each Constant-attribute value, making the last panel a valid inventory reference when inference is correct. Study II crosses both gates with each inference variant.

Candidate-side subscripts $j \in \{0, \ldots, 7\}$ below denote panel $8 + j$; the reference is $i = 7$. Define

$$w_{is} = \frac{\pi_{is}}{\sum_u \pi_{iu}}, \qquad H_{ia}(v) = \sum_s w_{is} q_{isa}(v). \tag{3}$$

Each candidate uses its own occupancy weights. The shared adapter clips occupancies to $[10^{-10}, 1 - 10^{-7}]$, ensuring positive pooling and presence denominators. Equation 3 is an occupancy-weighted average of heads, not the exact expectation of a normalized random inventory under uncertain occupancy.

For $a \in \mathcal{A}$, replace the native appearance comparison with

$$D_{Ha}(j) = \mathrm{JS}(H_{7a}, H_{ja}), \qquad S_H(j) = S_0(j) + \sum_{a \in \mathcal{A}} \left[D_{Ca}(j) - D_{Ha}(j) \right]. \tag{4}$$

All coefficients are one, with no target-answer fitting. When $\mathcal{A}$ is empty, $S_H = S_0$ exactly; other score terms and native choices are preserved.

Pooling costs $O(S\sum_a |\mathcal{V}_a|)$ per panel for S slots. Representation size is independent of occupied count, but the detector still assumes known nine-slot geometry; unseen-layout transfer does not follow.

Operational configuration. Final PrAE uses stable actual-mask inference (Section 3.4), the expanded gate and panel 7; controls retain original32 and stable-legacy branches. Study I uses original inference and gate. NVSA's primary uses native inference and averages panels 6 and 7. The all-Constant follow-up drops the Position gate. Study IV and its follow-up add scalar conditioning (Eq. 6); otherwise we use S_H. Answers take the first exact maximum; 10^{-6} tie-credit is diagnostic.

Presence and deletion controls. To separate distribution retention from simpler alternatives, let

$$e_{ia}(v) = \sum_s \pi_{is} q_{isa}(v), \qquad P_{ia}(v) = \frac{\min\{1, e_{ia}(v)\}}{\sum_u \min\{1, e_{ia}(u)\}}. \tag{5}$$

The presence control replaces D_{Ha} with $\mathrm{JS}(P_{7a}, P_{ja})$ in Eq. 4; deletion replaces it with zero. Both use the identical gate and other score terms. With exact one-hot objects, P is uniform over present values, while H retains their proportions. With uncertain heads, P is a clipped expected-count factor, inspired by NVSA's value-combination representation, and is not an exact presence posterior. When every expected count is at most one, $P = H$. These controls first enter the source study after Study I. They are then frozen before generating Study II, where all comparison choices within an inference variant share exactly the same inferred rules.

Ordinary-value comparison for scalar rules. For inferred non-Constant appearance actions, PrAE's executed law is $p_a = (\tilde{C}_a, 0)$: progression, arithmetic and D3 have no inconsistency output. Write a proper candidate law as $(r_{ja}c_{ja}, 1 - r_{ja})$, with $r_{ja} > 0$. The conditioning extension is

$$S_{HC}(j) = S_H(j) + \sum_{a \in \mathcal{N}} \left[D_{Ca}(j) - \mathrm{JS}(\tilde{C}_a, c_{ja}) \right], \tag{6}$$

where $\mathcal{N}$ contains inferred non-Constant actions and is disjoint from $\mathcal{A}$. Native divergence includes a consistency discrepancy and a category weight depending on r_{ja}; conditioning changes both. Consistency mass can help reject a heterogeneous distractor even when the inferred rule is correct. Conditioning trades that evidence for an ordinary-value comparison; its accuracy effect therefore requires testing. Conditioning executor inputs gives the same exact prediction, so no new executor is needed (Appendix Q). Study IV crosses candidate shape c versus H with retained versus unit ordinary mass, testing the benefit and harms at low counts.

3.3 Information retained beyond an inventory

With exact occupancy and deterministic heads, the comparison statistics form

$$F \longrightarrow J \longrightarrow H \longrightarrow P \longrightarrow C, \tag{7}$$

where F assigns object tuples to named slots, J retains the tuple distribution, H its attribute marginals, P their uniform value supports, and C within-panel consensus. Every arrow can discard information. Independent attribute matching permits a different partner for each attribute; joint matching keeps an object's attributes together; spatial matching also fixes the partner's location. For proper soft heads, the corresponding replacement costs obey $0 \le \sum_a D_{Ha} \le D_I \le D_J$. This orders costs, not answer accuracy; it does not order P costs. Appendix N defines the controls and Appendix A proves the relations.

3.4 A stronger controlled inference baseline

The structural gate supplies a sufficient inventory-preservation condition when the inferred rules are correct. Position Distribute Three tests this condition. The native position state at index i represents the nonempty bit mask $i + 1$. The generator-compatible three-state templates require equal object

counts and distinct masks. The legacy grouping admits indices $(1, 2, 4)$ because each has popcount one, although their actual masks $(2, 3, 5)$ have unequal object counts $(1, 2, 2)$. We evaluate the original support grouped by $\mathrm{popcount}(i)$ and the actual-mask support grouped by $\mathrm{popcount}(i + 1)$, keeping the remaining relation banks and native executors fixed.

Numerical evaluation is a separate factor. Products of context probabilities can underflow before normalization in float32; changing arithmetic can also change an answer that was previously correct. Our stable branches evaluate all rule evidence and normalizers in float64 from the same float32 scene log probabilities, retaining the native floors, template denominators and two-stage mode/rule selection. Position D3 evidence is summed with an established positive subset recurrence (Björklund et al., 2009). Scene laws, execution and the native candidate score stay in float32. Appendix F gives the recurrence and diagnostics. The support correction supplies a stronger baseline for testing the inventory contribution rather than attributing both effects to the comparison.

4 Why heterogeneous inventories become indistinguishable

Proposition 1 (Exact inventory collapse)**.** *Fix a nonempty occupied set and deterministic attribute heads. For an appearance attribute, $C = \delta_v$ if every object has value v, and $C = \delta_\perp$ otherwise. Thus any two heterogeneous inventories have the same native appearance law, whereas their histograms agree only when their normalized value counts agree.*

The proof follows directly from the product in Equation 1. If candidate Position, Number and remaining appearance summaries also agree, a scorer receiving only these summaries cannot recover the missing distinction. An index-based tie break can select one candidate but cannot supply evidence about its inventory. Figure 1 is an actual candidate pair with this complete ideal-summary collision.

The issue does not require a large count. Two objects with different Color values already enter $\perp$. More objects can intensify the contraction under specific heterogeneity conditions, but object count by itself is insufficient: an inventory with counts $(n - 1, 1)$ retains only a first-order exponent in the bound below.

Proposition 2 (Contraction for confidently correct heads)**.** *Assume exact occupancy of $n \geq 1$ objects and true categories t_s, with $q_s(t_s) \geq 1 - \epsilon$. Let $n_v = \sum_s \mathbf{1}[t_s = v]$. The total regular mass of the ideal native law satisfies*

$$r = \sum_v \prod_s q_s(v) \leq \delta := \min\left\{1, \sum_v \epsilon^{\,n - n_v}\right\}. \tag{8}$$

For two such panels with regular masses bounded by δ_P, δ_Q,

$$\mathrm{JS}(C_P, C_Q) \leq \max(\delta_P, \delta_Q) \ln 2. \tag{9}$$

If their true attribute histograms have disjoint supports and $0 \leq \epsilon \leq 1/2$, their soft histograms instead satisfy

$$\mathrm{JS}(H_P, H_Q) \geq \ln 2 - h_b(\epsilon), \tag{10}$$

where h_b is binary entropy in nats.

For a fixed heterogeneous inventory, every $n - n_v$ is positive. The native comparison therefore vanishes as the heads become correctly concentrated, while the histogram distinction can remain of constant order. If the query and correct candidate have the same ideal histogram, their soft histogram cost is at most $2\epsilon \ln 2$. A wrong candidate with disjoint ideal support then has a positive termwise cost margin whenever $\ln 2 - h_b(\epsilon) - 2\epsilon \ln 2 > 0$. Appendix B proves these statements.

The bounds identify a comparison margin under stated perception and rule assumptions. Other score terms and inference errors still determine the final answer. With arbitrary soft heads, histograms and native laws can retain different information: two equally uncertain binary heads and two opposing one-hot heads have the same histogram but different all-same probabilities. Appendix B gives the occupancy extension and counterexamples.

Table 1: **Distinct fresh cohorts, shared comparisons within each study.** Each study was frozen before generating its questions. Visual checkpoints are fixed across studies; Study III's two backends share visual0. The three PrAE branches separate original float32, stable legacy-index, and stable actual-mask inference. Opened development and the later reuse of Study I's high cohort are descriptive, not additional fresh evaluations.

Fresh study	Low	High	Visual models	Reasoning variants
I	1,100	2,200	3	Original PrAE; original gate
II	2,200	—	1	Three PrAE branches; both gates
III	1,100	2,200	1	Three PrAE branches and native NVSA
IV	3,300	—	1	Actual-mask PrAE; scalar-rule comparison

5 Experiments

5.1 Matched models, data and frozen comparisons

Setting. We generate count-controlled I-RAVEN questions with the pinned `raven-datasets` implementation (Małkiński & Mańdziuk, 2025), using a fixed nine-slot layout. For the local detectors in Studies I–III, training restricts every panel to one through four objects. Low-count evaluation applies this restriction to all sixteen panels; high-count evaluation requires at least one of the eight context panels to contain more than four objects. The final context (reference) panel need not have a high count. The public ALANS checkpoints' training ranges are undocumented, so their high-count evaluation tests visual-weight transfer, not verified unseen-count exposure.

The detectors were trained on 3,000 source questions and selected on 1,000 source-validation questions. Study I uses three fixed checkpoints with original PrAE inference. Studies II and III use visual0 and cross three PrAE inference branches with seven comparison variants. Study III additionally uses native NVSA inference and seven comparisons on the same visual posteriors. Every comparison within an inference branch shares its selected actions. Native PrAE scores are replayed exactly under the frozen adapter before intervention; neither perception nor planning is retrained.

Development and confirmation are distinct. Source validation and opened high-count cohorts establish the comparisons; their outcomes and later cohort reuses are descriptive. Each fresh study in Table 1 has a prospective freeze and eleven balanced Number/Position strata. ID, pixel, candidate-order-invariant pixel and XML/scene checks prevent overlap within the indexed scope (Appendices C, F.1 and J).

Study I endpoints. The high primary is H minus normalized-JS accuracy, paired by question and equally weighted across strata, with low-count noninferiority at a one-point margin. Its 10,000 shared bootstrap draws, non-sparsity conditions and tie sensitivity are specified in Appendix C. Subsequent studies freeze their own endpoints.

5.2 Study I: fresh count-range confirmation

Histograms raise high-count accuracy from 93.62% to 97.64%, a +4.02-point gain over normalized JS (95% paired interval [3.38, 4.70]), meeting the frozen criterion. Low-count accuracy rises from 93.79% to 95.24%: +1.45 [0.94, 2.03], meeting the one-point mean-loss guardrail. Table 6 retains all seven arms.

The three models correct 87/83/95 high and 17/15/16 low answers, with no harms. These are repeated model answers on shared questions.

Fractional credit among scores within 10^{-6} of the maximum preserves both signs: +4.20 high and +1.74 low points. Appendix E retains per-model, tie and stratum results.

Independent OT adds +0.02 high-count points over histograms; joint and spatial matching add +0.06, involving only one and two distinct questions. Their descriptive intervals [0.00, 0.05] and [0.00, 0.17] do not establish superiority. Spatial matching loses 0.12 low-count points (1/0/0 corrections, 2/1/2 harms). Marginal-value retention accounts for the dominant gain.

Table 2: **Study II: fresh corrected-inference confirmation.** One fixed visual0 on 2,200 new low-count questions. Core contrasts are shown here; Appendix F.1 retains every one of the 21 arms. These questions are distinct from Study I and from source development.

Inference and comparison	Correct	Accuracy (%)	Tie credit (%)
Original32 native	2063	93.77	93.58
Original32 old-gate H	2100	95.45	95.39
Stable legacy native	2055	93.41	93.22
Actual-mask native	2123	96.50	96.38
Actual-mask normalized JS	2123	96.50	96.38
Actual-mask old-gate H	2160	98.18	98.18
Actual-mask expanded H	2180	99.09	99.09
Actual-mask expanded presence	2180	99.09	99.09
Actual-mask expanded deletion	1989	90.41	90.48

Only the two Position Progression strata and Number/Position Constant improve: $+13.33$, $+15.17$ and $+15.67$ points. The other eight strata have unchanged H-versus-normalized-JS accuracy. Position Distribute Three remains at 76.17%, motivating the inference and expanded-gate comparison below.

The gain also occurs when reference panel 7 itself has more than four objects. On the 1,246 such questions, the three fixed models correct 52, 46 and 54 normalized-JS errors, with no harms. Appendix G reports this post-hoc count decomposition.

5.3 Study II: fresh confirmation with corrected inference

Study II tests the source decomposition of arithmetic, Position D3 support and comparison on new questions. Appendix F retains all 29 source arms and their five arithmetic harms. Twenty-one arms cross three inference branches and seven comparisons on 2,200 new low-count questions. Two frozen primaries test support correction and inventory comparison, using 20,000 paired within-stratum bootstrap draws and 97.5% intervals with sparse/degenerate safeguards. Original32 old-gate H supplies a separate one-point mean-loss guardrail.

Support correction adds +3.09 points (97.5% interval [2.41, 3.77]), with 68 corrections and no harms. Expanded inventory comparison adds +2.59 over corrected normalized JS ([1.86, 3.36]), with 57 corrections and no harms. Both meet their improvement criteria.

Of the 57 inventory corrections, 37 occur under the original gate and 20 on newly eligible Position D3 questions. Corrected inference improves native execution and exposes additional cases requiring attribute values.

Presence gives identical answers. Histograms correct 191 deletion errors with no harms. Precision alone causes 2 corrections and 10 harms relative to original32 native; the final combination causes 90 corrections and 10 harms relative to original32 old-gate H. The mean-loss guardrail passes despite these individual harms.

5.4 Study III: fresh transfer across reasoning backends

Both backends share visual0 but infer rules separately on 1,100 low and 2,200 high fresh questions, balanced across eleven strata. All 21 PrAE and seven NVSA arms were frozen before generation.

PrAE compares stable actual-mask expanded H with matched normalized JS at high counts and original32 original-gate H for the low guardrail. NVSA compares averaged H with normalized native execution in both ranges, sharing the panels 6–7 reference.

The 20,000 paired-stratified bootstrap draws resample questions jointly across arms; each of four claims has nominal $\alpha = 0.0125$. Low noninferiority allows one-point mean loss. Fractional tie credit at tolerance 10^{-6} preserves both high contrast signs; primary answers use literal first-index selection.

Table 3: **Study III: fresh confirmation in two backends.** Accuracy is comparator → H; changes/bounds are percentage points. C/H: corrections/harms. High: 97.5% paired intervals; low: 99.375% one-sided lower bounds L_B. Both high superiority criteria and low mean-loss guardrails pass. Original H denotes original32, original-gate inference.

Backend/range	Accuracy (%)	Δ	C/H	Bounds	Comparator
PrAE high	94.50 → 99.68	+5.18	114/0	$[+4.18, +6.23]$	Norm. JS
NVSA high	93.77 → 99.50	+5.73	126/0	$[+4.68, +6.82]$	Norm. native
PrAE low	94.73 → 99.09	+4.36	52/4	$L_B = +3.182$	Original H
NVSA low	95.91 → 98.82	+2.91	33/1	$L_B = +1.727$	Norm. native

Table 4: **Transfer to three published visual checkpoints.** All 2,200 high and 1,100 low questions are reused from Study III. High accuracy is matched baseline → H; C/H counts corrections/harms. Fixed means average checkpoints within each question; only these high rows show 97.5% paired-stratified intervals. Per-checkpoint and low effects are descriptive.

Backend / weights	High accuracy (%)	High Δ (pp)	C/H	Low Δ (pp)
PrAE / sys	76.77 → 81.23	+4.45	118/20	+2.64
PrAE / prod	78.23 → 83.41	+5.18	137/23	+3.18
PrAE / loc	76.23 → 80.82	+4.59	121/20	+3.09
PrAE / fixed mean	77.08 → 81.82	+4.74 $[+3.83, +5.65]$	—	+2.97
NVSA / sys	74.14 → 78.55	+4.41	112/15	+2.27
NVSA / prod	77.36 → 81.77	+4.41	112/15	+3.18
NVSA / loc	72.64 → 76.50	+3.86	100/15	+2.64
NVSA / fixed mean	74.71 → 78.94	+4.23 $[+3.41, +5.08]$	—	+2.70

PrAE's low guardrail is aggregate: all 52 corrections occur in Position D3; four arithmetic harms remain elsewhere, including three losses among 100 Number-addition questions. Neither low contrast has zero harms, and mean noninferiority does not establish stratum-wise safety.

Matched controls. PrAE support correction alone recovers 43 low and eight high answers without harms; expanded H adds 33 low and 114 high corrections over corrected normalized JS, also without harms. Expanded eligibility accounts for nine and 26, respectively. NVSA's reference-matched last-panel H/consensus comparison recovers 32 low and 127 high answers without harms.

H and reference-matched P agree on all low answers; at high counts H gains two answers without harms per backend. Against deletion, PrAE H recovers 88 low and 189 high answers; NVSA averaged H recovers 90 and 205, all without harms. These controls support retained attribute values as the main benefit.

Appendix K reports all arms, arithmetic/normalization controls, strata, subsets, tie intervals, both NI routes and eligibility checks. Secondary contrasts are descriptive.

5.5 Transfer to published visual checkpoints

We replace the local detector with each of the three released ALANS ObjectCNN states, retaining all 28 comparison arms on the opened Study III questions. The mapping and comparisons are fixed before this transfer evaluation. Every visual checkpoint improves in both backends (Table 4). The fixed-asset high-count gains are $+4.74$ points for PrAE and $+4.23$ for NVSA; both prespecified 97.5% interval lower bounds exceed one point. The gain therefore persists across independently produced visual weights, at substantially lower overall answer accuracy. Per-checkpoint harms remain in the table.

On high-count questions, H exceeds matched deletion by 7.05/6.79 points for PrAE/NVSA. Reference-matched H/P differences are smaller: $+0.35/+0.27$ points, descriptively. Low PrAE H/P answers coincide, while low NVSA H is 0.06 points below P. These controls again favor retaining attribute values without establishing universal frequency superiority. Appendix L specifies the transfer and all-arm records; the questions are reused and public-checkpoint training ranges are unknown.

5.6 Study IV: candidate conditioning for scalar appearance rules

We freeze Eq. 6 against S_H on 3,300 new low-count questions (300 per stratum, fixed visual0, actual-mask inference, 17 arms). Accuracy rises from 98.82% to 99.27%: 25 corrections, ten harms, $+0.45$ points with 95% interval $[+0.12, +0.82]$. The primary passes its non-sparsity and tie-credit criteria but misses the $+0.5$-point practical target.

The crossed mass/shape comparison is informative: C/H shapes give 3,261/3,262 correct answers at native mass and 3,276/3,283 at unit mass. Conditioning adds 15 or 21 answers; changing shape adds one or seven. Secondary H candidate-only has 22 corrections and no harms against S_H, but only 17 correctness discordances against conditioned C. Native64 alone changes no answer. Appendix Q retains all arms and strata; the frozen primary remains C conditioning.

Transfer of the conditioning extension. Reusing Study III's opened questions with all three public visual weights, C-output raises low/high means from 89.88/81.82% to 91.15/83.80%, improving each weight in both ranges. High-count changes are 236 corrections and 105 harms across overlapping checkpoint events. Native64 changes no answer; unit-mass H/P reach 83.89/83.62%, while matched deletion reaches 38.26%. These descriptive controls support the conditioning effect without a universal shape ranking (Appendix R).

Is the structural gate necessary? An eight-cell H/P $\times$ gate $\times$ conditioning ablation applies the same inventory comparison to every inferred Constant attribute. With H and conditioning, low/high public-weight means increase from 91.15/83.80% to 91.88/84.39%, with 22/145 harms alongside 46/184 corrections. The gate supplies a sufficient inventory-preservation condition, not an accuracy optimum (Appendix S).

5.7 Retained distinctions and changed answers

Study III's complete change census finds ideal-native collision, an eligible H distinction and correct-to-reference H equality in 147/147 PrAE and 158/159 NVSA primary corrections. Two high-count questions separate H from P by frequency (Appendix K.1). With public weights, these conditions plus a metadata-compatible gate hold in 200/376 PrAE and 213/324 NVSA high-count correction events, but also 3/63 and 1/45 harms. These overlapping events span 393 questions: retaining a distinction does not guarantee its successful use by soft scores (Appendix M).

The scalar extension has a separate complete census of 578 changes. Among public high-count corrections/harms, 495/505 and 198/205 active attribute terms match the recorded rules; all selected target and winner roles are homogeneous. Their soft inconsistency mass can therefore penalize uncertain correct values as well as wrong ones. Conditioning changes these penalties and categorical weights together, producing both improvements and harms (Appendix T).

6 Discussion and scope

The comparison statistic should retain the information constrained by the inferred rule. Inventory comparison implements this principle for Constant appearance: its gains persist beyond corrected PrAE inference, across backends and with public visual weights. Presence recovers most gains, while frequencies and bindings resolve additional ambiguities; matched deletion remains substantially worse.

Scalar conditioning adds fresh low-count and descriptive public-weight gains, with a larger mass than shape effect in the tested comparisons. The structural gate supplies a sufficient semantic condition, not an accuracy optimum. All studies use known nine-slot geometry, and public weights share a related CNN with undocumented training count ranges. Transfer to unseen layouts, clutter and other relation families remains open.

REPRODUCIBILITY STATEMENT

Frozen protocols and saved-result audits document all studies. The submitted supplement contains the public-weight image-input reconstruction kit, comparison source and complete scalar-comparison summaries. A local NumPy-only replay exactly reproduces 39,648 cached arrays and every decision/tie in Study IV and its public extension; its full cached inputs are not included in the size-limited supplement. This check reuses baseline anchors without rerunning perception, rule inference or bootstrap. A fresh installation remains untested. Appendix L maps supplied files to verified stages.

AI ASSISTANCE

Generative AI tools assisted implementation, experiment analysis, mathematical exposition, literature checking and manuscript drafting. The authors are responsible for verifying all claims, results and citations and for the final manuscript.

REFERENCES

Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. On evaluation of permanents. *arXiv preprint arXiv:0904.3251*, 2009. URL https://arxiv.org/abs/0904.3251.

Giacomo Camposampiero, Michael Hersche, Aleksandar Terzić, Roger Wattenhofer, Abu Sebastian, and Abbas Rahimi. Towards learning abductive reasoning using VSA distributed representations. In *Neural-Symbolic Learning and Reasoning*, volume 14979 of *Lecture Notes in Computer Science*, pp. 370–385. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-71167-1_20.

Michael Hersche, Mustafa Zeqiri, Luca Benini, Abu Sebastian, and Abbas Rahimi. A neuro-vector-symbolic architecture for solving raven's progressive matrices. *Nature Machine Intelligence*, 5:363–375, 2023. doi: 10.1038/s42256-023-00630-8. URL https://www.nature.com/articles/s42256-023-00630-8.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1567–1574, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16248.

Mikołaj Małkiński and Jacek Mańdziuk. A-I-RAVEN and I-RAVEN-Mesh: Two new benchmarks for abstract visual reasoning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5932–5940, 2025. doi: 10.24963/ijcai.2025/660. URL https://www.ijcai.org/proceedings/2025/660.

Shanka Subhra Mondal, Jonathan D. Cohen, and Taylor Whittington Webb. Slot abstractors: Toward scalable abstract visual reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36088–36105. PMLR, 2024. URL https://proceedings.mlr.press/v235/mondal24a.html.

Taylor Webb, Shanka Subhra Mondal, and Jonathan D. Cohen. Systematic visual reasoning through object-centric relational abstraction. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3151. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e3cdc587873dd1d00ac78f0c1f9aa60c-Abstract-Conference.html.

Yuhuai Wu, Honghua Dong, Roger Grosse, and Jimmy Ba. The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*, 2020. URL https://arxiv.org/abs/2007.04212.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual REasoNing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5317–5327, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_RAVEN_A_Dataset_for_Relational_and_Analogical_Visual_REasoNing_CVPR_2019_paper.html.

Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9736–9746, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Abstract_Spatial-Temporal_Reasoning_via_Probabilistic_Abduction_and_Execution_CVPR_2021_paper.html.

Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu. Learning algebraic representation for systematic generalization in abstract reasoning. In *Computer Vision – ECCV 2022*, volume 13699 of *Lecture Notes in Computer Science*, pp. 692–709. Springer, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-19842-7_40.

A Information retained by the comparison statistics

For a nonempty exact occupied set M of size n and deterministic heads $q_{sa} = \delta_{x_{sa}}$, let F be the assignment of tuples $x_s = (x_{sa})_{a\in\mathcal{A}}$ to named slots. Forgetting slot names produces $J = n^{-1}\sum_s \delta_{x_s}$, and coordinate marginalization produces $H_a = n^{-1}\sum_s \delta_{x_{sa}}$. Equation 5 then gives $P_a(v) = \mathbf{1}[H_a(v) > 0]/|\operatorname{supp}(H_a)|$. The native C_a is determined by P_a: it is δ_v for singleton support $\{v\}$, and $\delta_\perp$ otherwise. This proves Equation 7 and Proposition 1 under exact occupancy and deterministic heads, not for arbitrary soft heads.

Each reduction can be strict. Two inventories $(0, 1)$ and $(0, 2)$ give the same C but different P. Inventories $(0, 0, 1)$ and $(0, 1, 1)$ give the same P but different H. With two attributes, tuple sets $\{(0, 0), (1, 1)\}$ and $\{(0, 1), (1, 0)\}$ give the same marginals but different J. Swapping complete tuples between two named slots preserves J but changes F. The conclusions presume fixed Position/Number and remaining comparison inputs when used as score indistinguishability statements.

For arbitrary proper soft heads and any feasible coupling γ, joint convexity gives

$$\mathrm{JS}\left(\sum_s w_{7s} q_{7sa}, \sum_t w_{jt} q_{jta}\right) \le \sum_{s,t} \gamma_{st}\, \mathrm{JS}(q_{7sa}, q_{jta}).$$

Minimizing per attribute and summing proves $\sum_a D_{Ha} \le D_I$. A sum of separate minima cannot exceed the minimum of the sum, so $D_I \le D_J$.

With one-hot heads, local JS equals zero for equal values and $\ln 2$ otherwise. The maximum same-category matched mass is $\sum_v \min(H_{7a}(v), H_{ja}(v))$, yielding

$$D_I = (\ln 2) \sum_{a\in\mathcal{A}} \mathrm{TV}(H_{7a}, H_{ja}).$$

Thus independent matching vanishes exactly when all marginal histograms agree. Joint matching vanishes exactly when the weighted tuple distributions agree. Its ground cost is $\ln 2$ times the number of mismatching eligible attributes, rather than a binary whole-tuple mismatch. For soft heads, zero joint cost characterizes equality of empirical distributions of whole head vectors; it is not a statement about a latent joint posterior unavailable at the interface.

The spatial comparison uses $\gamma_{s,T(s)} = w_{7s}$. This is a feasible coupling for the OT controls only when $w_{j,T(s)} = w_{7s}$. Under that additional condition $D_J \le D_T$. Because the implemented spatial score does not enforce it, we do not impose an unconditional ordering involving D_T.

B Proof and limits of the soft-head bound

For category v, each slot with $t_s \neq v$ has $q_s(v) \le \epsilon$, while matching slots contribute at most one. There are $n - n_v$ mismatching slots, proving Equation 8. For native laws C_P, C_Q with regular masses r_P, r_Q,

$$\mathrm{TV}(C_P, C_Q) \le \tfrac{1}{2}(r_P + r_Q + |r_P - r_Q|) = \max(r_P, r_Q).$$

The categorical inequality $\mathrm{JS}(P, Q) \le (\ln 2)\,\mathrm{TV}(P, Q)$ then proves Equation 9. To see that inequality directly, write each coordinate pair with masses a, b as a JS contribution $(a+b)[\ln 2 - h_b(a/(a+b))]/2$. Concavity of binary entropy bounds it by $|a - b| \ln 2/2$, and sum over coordinates.

Let $H^* = n^{-1}\sum_s \delta_{t_s}$. Convexity of the ℓ_1 norm yields $\mathrm{TV}(H, H^*) \le \epsilon$. If the two ideal supports are disjoint, coarse-grain the alphabet to the support of H_P^* and its complement. The resulting Bernoulli probabilities obey $a \ge 1 - \epsilon$ and $b \le \epsilon$. Coarse-graining cannot increase JS, and binary JS is minimized on this rectangle at $(a, b) = (1 - \epsilon, \epsilon)$ for $\epsilon \le 1/2$. This gives Equation 10. Query and correct candidate with the same H^* have total variation at most 2ϵ; their JS is at most $2\epsilon \ln 2$.

If the largest category frequency is at most βn for fixed $\beta < 1$, then $r \le \min\{1, |\mathcal{V}|\epsilon^{(1-\beta)n}\}$. This is the condition under which the exponent increases with count. Heterogeneity alone at increasing n need not give that rate. At $\epsilon = 0$, interpret $0^0 = 1$: homogeneous inventories have regular mass one and heterogeneous inventories have regular mass zero. Empty occupancy is outside the setup, and a single object cannot form a heterogeneous inventory.

For uncertain occupancy with posterior mass at least $1-\eta$ on the true occupied set, mixture linearity gives the limited extension $r \le (1-\eta)\delta + \eta$. The actual histogram weights $\pi_s / \sum_u \pi_u$ still need a separate error analysis; that formula is not generally $\mathbb{E}[\mathbf{1}\{s \in M\}/|M| \mid M \neq \varnothing]$. For example, independent occupancies $(1, 1/2)$ yield weights $(2/3, 1/3)$, whereas the exact expected normalized inventory weights are $(3/4, 1/4)$. No empirical confidence or occupancy-concentration bound is inferred in this paper. Native probability floors and clamps also distinguish literal floating-point execution from these ideal distributions.

Histograms are not universally sufficient for uncertain native summaries. With two occupied objects and a binary attribute, heads $(1/2, 1/2), (1/2, 1/2)$ give native law $(1/4, 1/4, 1/2)$. Opposing one-hot heads give $(0, 0, 1)$. Their histograms coincide. The method is consequently a tested change of comparison statistic, not a lossless reparameterization for arbitrary soft heads.

C PROTOCOL, PROVENANCE AND TECHNICAL AMENDMENT

The published PrAE backend is pinned to revision `ca01d9c`. The I-RAVEN generator comes from `mikomel/raven-datasets`, revision `2c2630f`, distributed with A-I-RAVEN and I-RAVEN-Mesh (Małkiński & Mańdziuk, 2025). Its README identifies SRAN and I-RAVEN-Mesh as upstream implementations. We use its I-RAVEN branch with the count restrictions and nine-slot geometry specified here. Full revisions and file hashes are preserved in the protocol. The local probability adapter is separately hashed; Study I's three detectors are source-trained local models. The PrAE comparison uses the published reasoning backend with this shared perception interface; reported accuracies refer to that setup.

The adapter changes private import names, stores discrete Position transitions compactly, and evaluates long transition sums in chunks. It retains the author relation definitions, scene aggregation and scoring functions. Chunked floating-point reductions need not be bitwise identical to a dense implementation. Our exact replay claims compare the frozen adapter's full native inference with its cached-action execution in the same batch schedule.

Local perception supervision. Each local SlotDetector is freshly initialized with seed 0, 1 or 2 and uses the same 3,000 training and 1,000 validation questions. All sixteen panels contain one through four objects in the known nine-slot layout. Source XML supplies binary occupancy and Type, Size, Color and Angle labels for every context and candidate panel. The model uses 32×32 bilinearly resized grayscale slot crops, three convolutional layers with 32, 64 and 96 channels, and a 128-unit hidden layer, followed by one occupancy output and four ten-category heads. Its loss is mean occupancy binary cross-entropy over all slots plus mean attribute cross-entropy over the four heads at ground-truth occupied slots. The occupancy positive-class weight is the source negative/positive ratio clipped to $[0.25, 4]$, yielding 2.961776; both loss terms have unit coefficients. Empty-slot attributes are excluded, with no per-head weights or class-support masks. Angle participates in training and validation selection but is unused in reasoning. There is no answer or rule loss in local detector training.

Training uses 50 epochs of AdamW, batch size 1,024, learning rate 10^{-3}, weight decay 10^{-4} and gradient-norm clipping at 5. Source-validation occupied-attribute accuracy across all four heads, then occupancy accuracy and lower total loss determine checkpoint selection; selected epochs are 50, 48 and 49. All three checkpoints precede this study and stay fixed across comparisons. Their Type, Size and Color outputs pass through the same native-support projection for every arm. This local supervision differs from published PrAE's end-to-end answer and auxiliary rule supervision without visual attribute annotations. The public-weight study instead uses released ALANS ObjectCNN states, whose checkpoint-specific training count ranges are undocumented.

The adapter retains Type labels 1–5 and maps them to native indices 0–4, Size 0–5, and Color 0–9; Angle is unused in reasoning. Occupancies are clipped to $[10^{-10}, 1-10^{-7}]$. Selected categorical probabilities are floored at 10^{-10}, normalized in float32 and converted to log probabilities. The distribution-comparison arms exponentiate these stored log heads in float64 and normalize again. Earlier source examples use original XML labels; the later case censuses use native category indices as indicated. This shared projection is fixed, rather than inferred from fresh answers.

The scientific protocol froze before new generation. Low/high master seeds are 202609160320 and 202609160321; question IDs are 132000000–132003299. Both splits balance eleven Number/Position strata. Among completed raw questions, only the count-support predicate permits rejection along the fixed random stream; the pinned generator's internal construction and resampling remain in place. Schema, solver, duplicate or rendering failures stop and preserve outputs instead of selecting replacement seeds. The accepted questions require 1,216 low and 2,556 high raw attempts. All completed raw NPZ/XML artifacts are retained. The low and high Parquet conversions contain 17,600 and 35,200 panel images, respectively.

The historical index records 155,011 identities per identity definition across the inventoried canonical cohorts. It includes original training, source validation, opened development and subsequent indexed cohorts. It does not claim exhaustive current pixel decoding of every archived file on the computer; some historical pixel identities come from frozen manifests, and transformed rasters are pixel-covered only if separately indexed. Semantic scene identities cover the corresponding recorded scene. Checks also include candidate-order-invariant pixels and internal low/high duplication.

The first execution failed on Windows after writing the low Parquet file, because the validation reader still held its file handle at final rename. The original failed partial file, raw questions, log and protocol remain unchanged. A reviewed amendment closed that handle and independently revalidated all low rows and panel images before copying the identical Parquet bytes to the final path. High generation used the original seed and configuration. The amended implementation freeze occurred after low generation and before any model evaluation; it is not described as a wholly pre-data code freeze. Scientific choices, model files, scoring and analysis criteria were unchanged.

GPU perception is serial. Full native inference uses CPU batch size eight, one thread per worker, at most two workers, and transition chunk size 8,192. All six model/split score inventories must pass before answer-dependent analysis. Generation QA necessarily reads author labels to validate identity; model scoring uses no labels. The score barrier seals scores, native actions, tie sets, and LP certificates. The corrected pipeline has the original six-hour aggregate execution budget minus the elapsed failed attempt; per-stage and memory watchdogs preserve failure outputs.

For the high primary, the paired bootstrap interval must have positive lower bound, at least twenty nonzero question-mean differences, and nonzero variation in a stratum. Otherwise the prospective decision is inconclusive. Low noninferiority requires the paired lower bound to exceed -1 percentage point under the same regularity checks, or the prespecified zero-harm branch. With zero any-model harmful questions among N independent equally stratified questions, the one-sided 97.5% upper harm bound is $1 - 0.025^{1/N}$. Expected average-model loss is at most any-model harm probability. Equal allocation and the AM–GM bound on the product of no-harm probabilities permit differing stratum harm rates. This bounds mean loss, not future zero harm or each stratum's loss.

Bootstrap seeds are 202609160322 and 202609160323. Every draw carries all models and arms together. Reported intervals describe this fixed-model equal-stratum question distribution, not training-seed population uncertainty. Secondary pairwise intervals are unadjusted; their containing zero is not evidence of equivalence. Fractional tie credit assigns $1/k$ to a correct answer among k scores within 10^{-6} of the maximum. The integer denominator 840 represents all possible tie sizes from one through eight without rounding ambiguity.

An independently authored saved-statistics audit recomputed all six model/split outputs, all arm accuracies, 42 pairwise question contrasts, tie credits, strata and every one of the 10,000 paired bootstrap draws per range. The recomputation matched the saved results within the declared numerical tolerance. It imported no production scoring or analysis functions. A separately authored cohort audit checked all 3,300 accepted questions, all 3,772 completed raw attempts, and 52,800 raw-versus-Parquet panel image arrays, including counts, labels, rules, metadata and historical/internal identity separation. The root authoring agent executed both auditors. These checks independently implement saved-output validation; they are not a second pixel-to-logit model run or external scientific peer review.

D OPENED DEVELOPMENT AND SOURCE-CASE DETAILS

Spatial and joint matching correct 47, 41 and 43 normalized-JS errors, with no previously correct answer lost. Relative to the original native score, the corresponding counts are 47, 41

Table 5: All opened high-count model counts out of 1,000 shared questions.

Score	Visual 0	Visual 1	Visual 2
Native PrAE	931	939	936
Normalized JS	931	939	937
Attribute histogram	976	979	979
Unaligned slots	926	936	932
Spatial correspondence	978	980	980
Independent attribute OT	976	979	979
Joint object OT	978	980	980

Table 6: **Fresh confirmation.** Accuracy averaged over three fixed models on 1,100 low and 2,200 high questions. Questions, rather than model–question answers, are the sampling units. All seven frozen arms are shown.

Score	Low (%)	High (%)
Native PrAE	93.76	93.67
Normalized JS	93.79	93.62
Attribute histogram	95.24	97.64
Unaligned slots	90.82	93.92
Spatial correspondence	95.12	97.70
Independent attribute OT	95.24	97.65
Joint object OT	95.24	97.70

and 44 because normalized JS changes one visual-2 answer. The two additional visual-0 corrections over histograms concern questions 10004173 and 10004404. In the latter, the query's Type–Size tuples are $(5,3),(3,4),(3,0),(2,1),(1,0)$, while the incorrectly selected candidate has $(5,0),(3,1),(3,4),(2,0),(1,3)$. Marginals agree but their pairings differ. Joint and spatial comparison recover the correct candidate. These annotations explain fixed earlier predictions; they never select an inference gate.

The complete source correction set is 10003077, 10003118, 10003204, 10003274, 10003295, 10003321, 10003331, 10003362, 10003457, 10003604, 10003663, 10003756, 10003799, 10003897 and 10003951. The symbolic audit uses separately authored XML parsing and rational histogram arithmetic. It checks candidate Position and Number, noneligible appearance summaries, native selected actions, true rules and eligible marginal/joint distributions. All 15 meet the complete ideal-summary alias and query-histogram recovery conjunction. None has equal eligible histograms but different joint tuples. The selected set does not measure how often such cases occur in an unconditioned population.

Figure 1 uses source question 10003077, query panel 7, correct choice 3 and wrong choice 4. The raw 160×160 grayscale pixels agree with the saved Parquet images for all sixteen panels. Query slots $\{0,4\}$ move to $\{1,5\}$ under Progression $+1$. Color inventories are $\{0:1/2, 5:1/2\}$ for query/correct and $\{1:1/2, 4:1/2\}$ for wrong. The noneligible Size rule changes values, so query and candidate images need not look identical. Type and Color were eligible under the saved native gate.

E Fresh per-model and stratum results

The high primary has 133 nonzero question-mean contrasts; the low contrast has 31. The complete machine-readable result retains all 21 pairwise contrasts per range, including corrections, harms, answer changes, sparse/degenerate flags, paired intervals and tie sensitivities. The low-count any-model harm count is zero; its conservative one-sided 97.5% upper probability bound is 0.3348%. This is not a guarantee of future zero harm.

Table 7: Fresh per-model correct answers: low out of 1,100, high out of 2,200. The final columns give mean fractional tie-credit accuracy (%). Visual models 0, 1 and 2 share each question set.

	Low correct			High correct			Tie credit	
Score	V0	V1	V2	V0	V1	V2	Low	High
Native PrAE	1030	1031	1033	2061	2063	2058	93.47	93.39
Normalized JS	1031	1032	1032	2060	2061	2058	93.47	93.39
Attribute histogram	1048	1047	1048	2147	2144	2153	95.21	97.59
Unaligned slots	994	997	1006	2069	2063	2067	90.81	93.90
Spatial correspondence	1047	1046	1046	2148	2146	2154	95.10	97.65
Independent attribute OT	1048	1047	1048	2147	2144	2154	95.21	97.61
Joint object OT	1048	1047	1048	2148	2146	2154	95.21	97.65

Table 8: All eleven fresh strata. Histogram minus normalized-JS accuracy in percentage points; descriptive within-stratum estimates. Each stratum has 100 low and 200 high questions shared across models.

Stratum	Low change	High change
Number progression -1	0.00	0.00
Number progression +1	0.00	0.00
Position progression -1	5.00	13.33
Position progression +1	4.33	15.17
Number arithmetic -1	0.00	0.00
Number arithmetic +1	0.00	0.00
Position arithmetic -1	0.00	0.00
Position arithmetic +1	0.00	0.00
Number distribute three	0.00	0.00
Position distribute three	1.00	0.00
Number/Position constant	5.67	15.67

F SUPPORT-CORRECTED INFERENCE AND ITS SOURCE CONTROLS

For Position D3-left, the eight context roles are (A, B, C, B, C, A, C, A), so the template weights factor as $A_i = P_0(i)P_5(i)P_7(i)$, $B_j = P_1(j)P_3(j)$ and $C_k = P_2(k)P_4(k)P_6(k)$. D3-right uses (A, B, C, C, A, B, B, C). Evidence sums $A_iB_jC_k$ over distinct states within each equal-count group, divided by the native template count. Both D9 support conventions contain 5,136,096 templates, so changing support does not change that prior denominator.

The coefficient of xyz in $\prod_i(1 + A_ix + B_iy + C_iz)$ sums the injective assignments within a group. Keeping the eight subsets of $\{x, y, z\}$ and updating from the previous state prevents reuse of an index. This is the established transposed subset recurrence for commutative semirings (Björklund et al., 2009, Theorem 1(ii)), evaluated here in the log domain. It is linear in the number of position states for three roles, but there are $2^S - 1$ states for S slots; it is not polynomial in slot count. The recurrence is a numerical implementation of template marginalization, not a newly claimed permanent algorithm.

All-rule precision is controlled separately from support. The native row floor, rule formulas, template averaging, scale and two-stage selection remain fixed. A preserved first-batch float32 failure arose before answer analysis because a Type conditional was undefined. A separate eight-question probe found all seven float32 raw evidences equal to zero for one Type context, while native float64 evaluation of identical supplied logs gave a positive divided total 1.286×10^{-65}. All 24 question/attribute cells had finite float64 conditionals. This justified coherent evidence evaluation; it did not predict that every changed answer would improve.

The final eleven source errors and all five numerical harms are preserved with rules, actions, scores and margins. The harms span Number D3 (one), Number addition (one), Number increasing progression (two) and Position union (one); every case changes a used appearance-rule action. This is a descriptive localization, not an isolated causal intervention on that appearance action. No exception or subgroup

Table 9: Opened-source decomposition on the same 1,000 questions and visual0. Accuracy differences are descriptive; corrections and harms use paired first-index answers.

Later arm versus earlier arm	Corrected	Harmed	Change (pp)
Stable legacy native / original native	2	5	−0.30
Actual-mask native / stable legacy native	30	0	+3.00
Actual-mask normalized JS / native	1	0	+0.10
Actual-mask old-gate H / normalized JS	15	0	+1.50
Actual-mask expanded H / old-gate H	8	0	+0.80
Actual-mask expanded H / normalized JS	23	0	+2.30
Actual-mask expanded H / deletion	89	0	+8.90
Actual-mask expanded H / saved expanded H	36	5	+3.10

filter is fitted to these cases. In contrast, H and presence make identical source decisions despite differing scores, so frequency necessity remains unresolved.

The source64 run took 142.482 seconds with observed peak RSS 879,861,760 bytes. It included two stable support branches, every native-action score replay, recomposition and full diagnostics; it was not a matched speed benchmark against complete original abduction. A separate saved-output auditor checked 130,821 assertions and 2,030 pair summaries. Its initial metadata parser omitted the combined Number/Position Constant alias; the preserved repair changed the auditor, not the experiment.

F.1 Complete fresh Study II results and protocol

Table 10: All 21 fresh Study II correct-answer counts, each out of the same 2,200 questions. Every column shares visual0 perception.

Comparison	Original32	Stable legacy	Stable actual mask
Native	2063	2055	2123
Normalized JS	2063	2055	2123
Old-gate histogram	2100	2092	2160
Old-gate presence	2100	2092	2160
Expanded histogram	2109	2101	2180
Expanded presence	2109	2101	2180
Expanded deletion	1953	1945	1989

Table 11: All eleven fresh Study II strata, 200 questions each, under actual-mask stable inference. JS denotes normalized native JS; H and P use the expanded gate.

Structural rule	JS	H	P	Deletion
Number Progression -1	200	200	200	200
Number Progression 1	199	199	199	199
Position Progression -1	188	200	200	156
Position Progression 1	187	200	200	146
Number Arithmetic -1	199	199	199	199
Number Arithmetic 1	196	196	196	196
Position Arithmetic -1	200	200	200	200
Position Arithmetic 1	191	191	191	191
Number Distribute Three 0	197	197	197	197
Position Distribute Three 0	178	198	198	148
Number/Position Constant 0	188	200	200	157

The new IDs are 135000000–135002199, with master seed 202609170350. Generating the 2,200 accepted questions required 2424 raw attempts. All 35,200 persisted panel images passed conversion checks. The exclusion index contains 158,797 identities under each of five definitions, including all

prior Study I attempts and fourteen subsequently constructed candidate variants. An initial metadata scanner failed on a historical Python 2 file; the preserved lexical-integer repair completed before any new generation or model evaluation. The full scientific and executable protocol was then frozen before drawing this cohort.

Tie-credit primary intervals are [2.48, 3.84] points for support correction and [1.99, 3.49] for inventory comparison. The binary primary contrasts contain 68 and 57 nonzero questions, respectively. The saved record retains all 210 full-cohort pairwise differences and all eleven stratum count tables. A separately authored statistical auditor recomputed winners, ties, exact counts, all 2,520 pair/subset comparisons, and the 20,000 bootstrap draws. Its author also implemented combined scoring, so its independence concerns statistical analysis; it is not an independent reimplementation of the complete vision pipeline. No empirical model or rule inference was rerun for this audit.

G POST-HOC MECHANISM AND REFERENCE-COUNT CENSUSES

These descriptive analyses use saved decisions, inferred actions and exact author XML annotations after both studies were opened. No model, scorer, rule inference, generation or hypothesis test was rerun. Their interpretation is separate from the frozen accuracy endpoints.

Complete Study II correction set. Expanded histogram comparison corrects 57 normalized-JS errors and harms none under actual-mask stable inference. The 37 original-gate corrections and 20 new D3 corrections are disjoint and exhaust this set. All 57 wrong/true candidate pairs have identical complete ideal native summaries. Eligible histograms distinguish 56 pairs; in all 57, the correct eligible histogram equals the query and every eligible true appearance rule is Constant. The saved record retains all case identities, rational histograms, rule annotations and failed conjunctions.

The exception is question 135001315 (Position Progression $+1$): the correct candidate is 1 and the normalized-JS answer is 6, using zero-based candidate indices. Type and Size are eligible. Query, correct and wrong candidates all have Type inventory $\{0 : 1/2, 1 : 1/2\}$ and Size inventory $\{3 : 1/2, 5 : 1/2\}$ in native category indices. Both candidates also have equal complete ideal native summaries. The changed soft-score answer cannot be attributed to recovery of a distinction in these exact eligible histograms. This census does not determine its cause.

Across all 2,200 questions, the original gate activates on 393 and the expanded gate on 527, including 134 newly active D3 questions. There are 693 eligible attribute cells and 231 questions with a heterogeneous eligible query attribute. Every active question satisfies the annotation-based sufficient condition: an inventory-preserving structural family and Constant eligible appearance rules. This checks compatibility with the invariance condition, not exact correctness of every selected rule action or a guarantee for future questions.

Study I high-count queries. All 2,200 questions have context maximum count between five and nine. Reference panel 7 itself has more than four objects on 1,246 questions and at most four on 954. Table 12 gives the complete count decomposition. The same questions are shared across models; the 265 corrected model answers concern 133 distinct questions. These results show gains on high-count and low-count reference panels within the high-context cohort. They do not establish a monotonic count effect or evaluate Study II's corrected pipeline at high counts.

Each fixed model has 374 original-gate eligible questions, including 244 with a heterogeneous eligible query. The source XML counts and rules, manifest identities, saved actions, reconstructed gates and candidate outcomes were checked across all questions. Exact query heterogeneity is an annotation property; its absence does not exclude effects from uncertain perception.

H NVSA: MATCHED REFERENCES AND NATIVE OUTPUT MASS

The NVSA application is pinned to revision `060425f` and operator library to `5499e1b`. A NumPy generator seeded at 1234 constructs the continuous and discrete codebooks in the original order, using dimension 1024 and four blocks. The codebooks remain fixed. The detector and its source-validation checkpoint selection are the same as in the preceding source studies. This census has no independent frontend or newly generated holdout.

Table 12: Post-hoc Study I decomposition by the exact number of objects in reference panel 7. Corrections compare histogram with normalized JS under original inference and the original gate. All harm counts are zero.

Reference objects	Questions	Visual 0 corrections	Visual 1	Visual 2
1	238	0	0	0
2	267	8	10	8
3	246	14	15	18
4	203	13	12	15
5	312	20	12	16
6	331	11	10	13
7	320	13	17	17
8	238	2	2	2
9	45	6	5	6
≤ 4	954	35	37	41
> 4	1246	52	46	54
All	2200	87	83	95

The native scene engine, relation bank, two-stage Position/Number selection and appearance execution are retained. A compact Position executor avoids dense cubic storage; source-free contracts compare it against dense four-slot execution and independently specified nine-slot semantics. A disclosed near-tie winner difference under the established numerical tolerance means this is not a claim of bitwise equivalence to every dense nine-slot operation. Native D3 support and normalization remain unchanged; the PrAE support correction is not applied here.

For an eligible Constant attribute, normalized native comparison uses the normalized sum of the scene laws of panels 6 and 7. The last-panel consensus control instead uses C_7, and the two histogram arms use H_7 or $(H_6 + H_7)/2$. All seven arms share the native actions. The six normalized arms share the gate and remaining score terms; original native scoring is retained separately. Presence uses Equation 5. All coefficients are one.

Table 13: Complete NVSA opened-source census, one fixed shared perception model and 1,000 questions. There is no new-data confidence interval or significance claim.

Comparison	Correct	Accuracy (%)	Tie credit (%)
Native score	971	97.10	97.05
Normalized native	973	97.30	97.05
Last-panel consensus	973	97.30	97.05
Last-panel histogram	993	99.30	99.30
Averaged histogram	993	99.30	99.30
Last-panel presence	993	99.30	99.30
Eligible-term deletion	904	90.40	90.46

H corrects twenty normalized-native/last-panel-consensus errors without harms. It corrects 89 deletion errors without harms. Relative to normalized native, deletion corrects one question, harms seventy and changes one wrong answer to another. Normalization alone corrects two literal native answers; its tie-credit change is zero. H, averaged H and presence agree on every answer, so these data establish neither a frequency-specific benefit nor superiority of one reference convention. The gate activates on 232 questions: 127 eligible Type, 86 Size and 86 Color cells. All 21 pairwise summaries are retained in the experiment record.

Complete candidate and gate census. A post-hoc audit joins all 1,000 saved predictions to the original, hash-verified source annotations. In each of the twenty corrections, the correct candidate and former wrong candidate have identical complete ideal Position, Number, Type, Size and Color summaries. Their eligible histograms differ, and the correct histogram equals the query histogram. The corrections divide into five predicted Constant, eight Progression and seven Distribute Three cases. All 232 active gates agree with a true Constant appearance rule and an inventory-preserving

structural family; 93 gated queries have heterogeneous eligible attributes. This agreement supports eligibility, not correctness of every selected structural action. In particular, metadata that does not encode Distribute Three orientation cannot verify that orientation. The census documents a compatible mechanism rather than proving the exclusive cause of the soft-score changes. Presence equivalence continues to leave frequency necessity unresolved.

Native executor output mass. The first attempted census stopped before labels because three native Number predictions had total mass 0.17303732. The original assertion required unit mass, but the published Identity decoder calls L1 normalization with denominator $\max(\|x\|_1, 10^{-12})$. It can therefore return positive subunit weights. The failed run is preserved. A separate protocol uniformly permits finite, nonnegative executor weights of positive finite total, retaining their exact values in native scoring; zero or invalid mass still stops. Scene laws, conditionals and actual selector inputs retain normalized-probability checks before selection. The existing normalized-native arm separately converts positive weights to proper distributions; its definition did not change.

In the complete census, 143 Number predictions lie outside unit-mass tolerance, all in inactive Number branches. One Type prediction also has subunit mass and contributes to scoring. The other three fields are within tolerance. Thus the amended domain is not an inactive-branch exemption. The native score on subunit weights is its original arithmetic expression, not a proper-distribution JS divergence.

The completed pipeline took 101.078 seconds on one CPU, with observed peak process-tree RSS 662,880,256 bytes; no detector forward or training ran. After all 125 batches and seven arms were sealed, a separate process read labels. An independently written NumPy auditor reconstructed scene-to-score arithmetic, selector outputs, gates, inventories, all score arms, ties and paired counts from saved artifacts. It imported no Torch, NVSA or producer code and took 4.593 seconds. Counts use literal saved argmax; tolerance-based cross-library checks never remap winners. This validates the saved computation, not an independent end-to-end replication.

I OPENED HIGH-COUNT FOLLOW-UP

This census uses all 2,200 previously evaluated Study I high questions with fixed visual0 posteriors and eleven original strata of 200. It repeats neither perception nor original full abduction. Original selected actions are replayed against all 275 archived executor and scene batches; both stable branches genuinely infer their own actions. The 21 new comparisons and seven archived comparisons are retained as distinct score banks. A complete score barrier precedes a separate label-analysis process, but the historical population remains opened. No new-data confidence interval or significance claim is assigned.

Table 14: All opened-high correct-answer counts out of 2,200, one shared visual0. Inference and comparison factors are crossed.

Comparison	Original32	Stable legacy	Actual mask
Native	2061	2060	2066
Normalized JS	2060	2059	2066
Original-gate histogram	2147	2146	2153
Original-gate presence	2145	2144	2151
Expanded histogram	2182	2181	2193
Expanded presence	2179	2178	2190
Expanded deletion	2006	2005	2009

The separate archived counts are native 2,061; normalized JS 2,060; histogram 2,147; unaligned 2,069; spatial correspondence 2,148; independent attribute OT 2,147; joint OT 2,148. These last four controls retain original inference; they are not matched comparisons for the stable branches.

Corrections, harms and ties. Precision alone produces zero corrections and one harm. Actual-mask support versus stable legacy yields six corrections and no harms (+0.27 points). Under actual-mask inference, expanded histogram versus normalized JS yields 127 corrections and no harms (+5.77

points; fractional tie-credit change +6.06). Expanding the gate adds forty corrections without harms over the original gate. Histogram exceeds expanded presence by three corrections without harms (+0.14 points), and deletion by 184 corrections without harms (+8.36 points). Presence itself corrects 125 normalized-baseline errors and harms one: H's three advantages comprise two additional baseline-error corrections and avoidance of that one harm. The three-question presence contrast is too sparse to establish a broad frequency-specific advantage.

Against archived original-gate histogram, the complete method corrects 47 questions and harms one (+2.09 points). The 1,800-question complement of both Number and Position D3 loses one answer (-0.056 points); the 2,000-question complement of Position D3 also loses one (-0.050 points). Both remain within the predeclared one-point practical screen. This is neither an individual no-harm result nor an opened-data noninferiority proof.

Table 15: Every original high-count stratum under actual-mask inference. Counts are out of 200 per row. H/P/Del denote expanded histogram, presence and deletion.

Stratum	Normalized	H	P	Del	H correction/harm
Number Prog. -1	200	200	200	200	0/0
Number Prog. $+1$	199	199	199	199	0/0
Position Prog. -1	174	200	200	154	26/0
Position Prog. $+1$	169	200	199	149	31/0
Number subtraction	200	200	200	200	0/0
Number addition	197	197	197	197	0/0
Position difference	200	200	200	200	0/0
Position union	200	200	200	200	0/0
Number D3	200	200	200	200	0/0
Position D3	158	198	197	153	40/0
Number/Position Constant	169	199	198	157	30/0

All three histogram–presence disagreements. A saved-only census covers the complete 2,200-question denominator and every H/P disagreement, using exact rational inventories reconstructed from the original XML. There are precisely three disagreements, all correct under H and wrong under P. In every pair, the correct candidate and the P-selected distractor have identical ideal Position, Number and all three native appearance summaries. Their ideal presence laws also agree on every eligible attribute, while their eligible histograms differ; the correct histograms equal the query. Table 16 displays the differing multiplicities. These are actual frequency-sensitive representational witnesses. They do not establish a general performance advantage or prove that frequency alone caused each uncertain-perception score change.

Table 16: Every opened-high H/P disagreement. Value orders use native category indices. Entries are exact counts; normalize by their sum to obtain H. P is uniform on the same listed values for both candidates. Query and correct-candidate counts agree.

Question	Attribute	Value order	Query/correct	P-selected
132001217	Type	$(0,1,2,3,4)$	$(1,2,2,1,1)$	$(1,1,2,2,1)$
132001630	Size	$(0,1,2,3,5)$	$(2,1,3,1,1)$	$(2,1,1,3,1)$
132003122	Type	$(0,1,2,3)$	$(1,1,1,3)$	$(2,1,2,1)$

For the correct-versus-P-selected candidate pair, H/P score margins are respectively $0.022319/-0.001951$, $0.057047/-0.008302$ and $0.071785/-0.000117$. The first margin change occurs in Type and the second in Size. The third is dominated by Type, with a 5.33×10^{-7} Color contribution despite equal ideal eligible Color inventories. Question 132001630 becomes correct when the D3 gate is expanded without changing its inferred actions. Question 132003122 is already correct under normalized consensus and H; P introduces its error. Thus the three H/P differences are not three additional normalized-baseline errors corrected by H. The full record preserves the soft costs, all 28 answers and these exceptions.

The complete method's one archival harm. Question 132001964 has true Number addition and a one-object query within a high-count context. All branches select Number mode and all inventory gates are off. The original branch answers candidate 5 correctly; both stable branches answer candidate 1. Size and Color selections change from Constant to addition at the precision transition. All seven original and all seven archived comparisons are correct, while all fourteen stable comparisons are wrong. Actual-mask support and inventory replacement make no further answer change in this case. This locates the loss in the stable inference transition without isolating an exclusive Size-versus-Color cause. All candidate indices here are zero-based.

The census used saved predictions and annotations only (0.250 seconds, one CPU, approximately 70 MB peak RSS). It is post-hoc interpretation by the implementation author, distinct from the independent saved-output audit.

Inference and archive integrity. Original-gate eligibility is unchanged across the three branches (374 questions). Expanded eligibility is 481 under original and stable-legacy inference and 499 under actual-mask inference. Corrected support changes 25 structural mode choices from Number to Position and one Position action; eighteen questions become newly eligible. No gate was selected from correctness or annotations.

All original native scores, selected actions, executed predictions and scene laws replay exactly. Recomputed normalized/histogram score bridges differ from historical values by at most 4.44×10^{-16}, with zero literal winner or tie-set differences. Archived outcomes are still retained separately. The original diagnostic nonfinite probabilities remain recorded; both new stable branches satisfy finite checks before selection.

The complete score and analysis run took 391.953 seconds on one CPU, followed by a 9.609-second independently implemented saved-output audit. The latter checked all 275 batches, thirty declared subsets and 7,170 pair summaries, including normalized components, projected-log inventories, gates, ties and archived linkage. It ran no model, planner or executor. All positive, equal and adverse results remain in the saved experiment record.

J STUDY III: PROSPECTIVE TWO-BACKEND EXECUTION

Study III tests the already specified PrAE and NVSA comparisons on the same newly generated images. The protocol, scientific source, dependencies, detector, codebook configuration, analysis and source-free checks were pinned before generation. The 236-artifact freeze has SHA-256 `383ad5a8f94830ef16cb55d98d2868221c28f5e2abc88546e9e182bb305f79bd`. Low/high generation seeds are 202609174110 and 202609174111. IDs 141000000–141001099 identify the 1,100 low questions; IDs 141001100–141003299 identify the 2,200 high questions. Each of eleven strata contributes 100 low and 200 high questions. These questions are distinct from the preceding studies within the indexed identity scope.

The historical index covers 161,221 identities per content definition, including all 2,424 raw Study II questions, not only its accepted low-count questions. The five definitions cover ordered raster, candidate-order-invariant raster, raw XML, semantic scene, and candidate-order-invariant semantic scene content. A separate namespace scan records 140,006 question IDs. High generation also excludes all completed raw low questions. Only failure of the prespecified count-support predicate permits rejection along the fixed generator stream. Duplicate, schema, solver, rendering and exhausted-budget failures stop the run; they do not trigger replacement seeds. The original generator's internal sampling remains unchanged.

Both backends receive the same fixed visual0 posterior arrays. PrAE selects its own original float32 actions on every new question; cached execution must reproduce those native scores exactly. Its stable legacy and stable actual-mask branches separately select their actions from the same posterior. Each branch then supplies all seven comparison arms. NVSA selects its own actions with codebook seed 1234, dimension 1,024 and four blocks; it does not use PrAE's support repair or actions. All five native executor fields must be finite, nonnegative and have strictly positive total weight. Positive subunit totals remain unchanged in the raw native arm, with renormalization represented by a separate control. Undefined or zero-total fields cannot be replaced by a favorable fallback.

The protocol requires every backend/range partition to have either complete, hash-verified scores for all its questions and arms, or an immutable failure receipt, before answer-dependent accuracy analysis. Scoring receives images or posteriors and question IDs, not answer labels or rule annotations. Generation quality control reads annotations to verify identity and count predicates. A local numerical failure preserves the entire failed partition and forbids partial accuracy; a shared-data, identity, supervision or resource failure stops dependent work. There are no automatic scientific retries.

Execution is serial, with one CPU thread per symbolic worker and fixed-checkpoint GPU perception. The overall cap is four hours, with a 12-GiB new-output cap, sampled recursive resident-memory cap of 4.5 GiB per worker and 2-GiB host-memory reserve. Full native abduction, stable comparison, generation, analysis and independent saved-output audit have separate stage limits within that shared budget. Stage receipts describe different computational tasks and are not matched speed comparisons between backends.

The separately authored NumPy auditor recomposes controlled comparison scores, checks recorded actions and native replay lineage, and reconstructs winners, tie credit, every paired contrast and the prespecified statistical decisions. It treats PrAE's recorded native executor outputs as inputs, rather than independently re-executing those operations. It does not rerun the detector or prove equality to an author's released end-to-end checkpoint. The subsequent saved-case census was specified before outcomes were opened: all corrections and harms for both primary inventory contrasts and the NVSA last-panel histogram/consensus contrast, all PrAE and reference-matched NVSA histogram/presence answer disagreements, and every low-count PrAE guardrail harm. It is an event-conditioned explanatory follow-up, not another confirmatory accuracy test or an estimate of collision prevalence among unselected questions.

K STUDY III: COMPLETE COMPARISONS AND PRESPECIFIED SUBSETS

The two backends share the newly generated questions, fixed visual0 perception and known nine-slot geometry. They are not independent frontend replications. PrAE's original32 actions are freshly abducted; stable legacy and actual-mask branches change inference separately from comparison. NVSA retains its native relation choices and seed-1234 codebooks (dimension 1024, four blocks). Low requires all sixteen panels to have at most four objects; high requires at least one context panel to exceed four. Correct-completion counts refer to panel $8 + y$, where y is the zero-based correct answer. These differ from the panel-7 reference counts in Study I.

Frozen analysis. Questions are resampled within the eleven strata, carrying all completed backend/arm outcomes jointly. The 20,000 draws use seeds 202609174112 (low) and 202609174113 (high), with linear percentiles. High primaries use two-sided 97.5% intervals and require a positive lower endpoint, at least twenty nonzero paired correctness differences, positive within-stratum variation, nondegenerate resampling and no fractional-tie sign reversal. The four affirmative claims each receive nominal $\alpha = 0.0125$. Each low NI claim uses two $\alpha = 0.00625$ routes: an eligible bootstrap lower bound L_B above -1 point, or a zero-observed-harm upper bound U_0 below one point. These are one-sided 99.375% nominal bounds. L_B is the 0.625th percentile of the saved bootstrap differences; with zero harms among n questions, $U_0 = 100(1 - 0.00625^{1/n})$ percentage points. The latter route is unavailable after any observed harm. Neither route establishes zero individual risk; bootstrap coverage is approximate. Decisions use unrounded values and all frozen eligibility checks.

References and controls. PrAE uses the final context panel for H/P comparisons. NVSA's primary averages H from panels 6 and 7, preserving its native averaged reference; last-panel H versus consensus is the matched-reference corroboration and last-panel H versus P is the frequency contrast. Raw native scoring remains a separate arm. H is the occupancy-weighted histogram; P is the normalized clipped expected-count control, not an exact presence posterior. All weights and gates are fixed.

The frozen decisions are PrAE (high superiority: met; low mean-loss guardrail: passed); NVSA (high superiority: met; low mean-loss guardrail: passed).

All four backend/range partitions completed and passed the saved-output audit.

Guardrail routes and tie sensitivity. PrAE low: bootstrap lower bound $+3.18182$ points; bootstrap route passes; zero-harm upper bound unavailable; zero-harm route does not pass. Observed corrections/harms are 52/4.

PrAE high: fractional tie-credit change $+5.39$ points, 97.5% interval $[+4.38, +6.42]$; no sign reversal. The literal contrast has 114 nonzero paired differences; bootstrap eligibility is satisfied.

NVSA low: bootstrap lower bound $+1.72727$ points; bootstrap route passes; zero-harm upper bound unavailable; zero-harm route does not pass. Observed corrections/harms are 33/1.

NVSA high: fractional tie-credit change $+5.93$ points, 97.5% interval $[+4.88, +7.00]$; no sign reversal. The literal contrast has 126 nonzero paired differences; bootstrap eligibility is satisfied.

Table 17: Study III: all PrAE comparison arms. Low and high ranges contain 1,100 and 2,200 questions, respectively. Tie denotes fractional tie-credit accuracy. Dashes cover an entire failed partition; no partial accuracy is reported.

	Low			High		
Comparison	Correct	Acc. (%)	Tie (%)	Correct	Acc. (%)	Tie (%)
Original float32						
Raw native	1018	92.55	92.48	2066	93.91	93.78
Normalized JS	1018	92.55	92.48	2067	93.95	93.78
Original-gate H	1042	94.73	94.73	2156	98.00	97.97
Original-gate P	1042	94.73	94.73	2154	97.91	97.88
Expanded-gate H	1045	95.00	95.00	2177	98.95	98.97
Expanded-gate P	1045	95.00	95.00	2175	98.86	98.87
Expanded-gate deletion	976	88.73	88.60	1995	90.68	90.60
Stable legacy support						
Raw native	1014	92.18	92.11	2070	94.09	93.96
Normalized JS	1014	92.18	92.11	2071	94.14	93.96
Original-gate H	1038	94.36	94.36	2159	98.14	98.11
Original-gate P	1038	94.36	94.36	2157	98.05	98.02
Expanded-gate H	1041	94.64	94.64	2180	99.09	99.10
Expanded-gate P	1041	94.64	94.64	2178	99.00	99.01
Expanded-gate deletion	972	88.36	88.23	1999	90.86	90.78
Stable actual-mask support						
Raw native	1057	96.09	96.02	2078	94.45	94.29
Normalized JS	1057	96.09	96.02	2079	94.50	94.29
Original-gate H	1081	98.27	98.27	2167	98.50	98.44
Original-gate P	1081	98.27	98.27	2165	98.41	98.35
Expanded-gate H	1090	99.09	99.09	2193	99.68	99.68
Expanded-gate P	1090	99.09	99.09	2191	99.59	99.59
Expanded-gate deletion	1002	91.09	90.78	2004	91.09	90.99

Table 18: Study III: all NVSA comparison arms. Low and high ranges contain 1,100 and 2,200 questions, respectively. Tie denotes fractional tie-credit accuracy. Dashes cover an entire failed partition; no partial accuracy is reported.

	Low			High		
Comparison	Correct	Acc. (%)	Tie (%)	Correct	Acc. (%)	Tie (%)
Raw native	1055	95.91	95.84	2061	93.68	93.57
Normalized native	1055	95.91	95.84	2063	93.77	93.57
Last-panel consensus	1056	96.00	95.93	2062	93.73	93.54
Last-panel H	1088	98.91	98.91	2189	99.50	99.50
Averaged H	1087	98.82	98.82	2189	99.50	99.50
Last-panel P	1088	98.91	98.91	2187	99.41	99.41
Deletion	997	90.64	90.64	1984	90.18	90.15

The following tables retain every structural stratum and the declared correct-completion-count and D3 exclusions. Low PrAE is shown both against its deployment guardrail and against matched

normalized JS; the latter is descriptive at low counts. No subset replaces a full-cohort primary or guardrail. All 210 PrAE and 21 NVSA within-backend pairs per complete range, including their stratum/subset summaries, remain in the machine-readable analysis record bound by the table-lineage receipt.

Table 19: Study III: PrAE low guardrail. Base is original32 original-gate H; H is expanded H with stable actual-mask inference for PrAE and averaged H for NVSA. Changes are in percentage points; C/H denotes corrections/harms. Strata and subsets are descriptive, with no separate inferential claim. An empty completion-count subset has $n = 0$.

Stratum / subset	n	Base (%)	H (%)	Δ	C/H	Tie Δ
Number progression -1	100	100.00	100.00	+0.00	0/0	+0.00
Number progression $+1$	100	100.00	100.00	+0.00	0/0	+0.00
Position progression -1	100	100.00	100.00	+0.00	0/0	+0.00
Position progression $+1$	100	100.00	100.00	+0.00	0/0	+0.00
Number arithmetic -1	100	99.00	98.00	-1.00	0/1	-1.00
Number arithmetic $+1$	100	98.00	95.00	-3.00	0/3	-3.00
Position arithmetic -1	100	100.00	100.00	+0.00	0/0	+0.00
Position arithmetic $+1$	100	99.00	99.00	+0.00	0/0	+0.00
Number Distribute Three	100	100.00	100.00	+0.00	0/0	+0.00
Position Distribute Three	100	48.00	100.00	+52.00	52/0	+52.00
Number/Position Constant	100	98.00	98.00	+0.00	0/0	+0.00
Completion count ≤ 4	1100	94.73	99.09	+4.36	52/4	+4.36
Completion count > 4	0	—	—	—	—	—
Excluding both D3 strata	900	99.33	98.89	-0.44	0/4	-0.44
Excluding Position D3	1000	99.40	99.00	-0.40	0/4	-0.40

Table 20: Study III: PrAE high primary. Base is matched normalized JS; H is expanded H with stable actual-mask inference for PrAE and averaged H for NVSA. Changes are in percentage points; C/H denotes corrections/harms. Strata and subsets are descriptive, with no separate inferential claim. An empty completion-count subset has $n = 0$.

Stratum / subset	n	Base (%)	H (%)	Δ	C/H	Tie Δ
Number progression -1	200	100.00	100.00	+0.00	0/0	+0.00
Number progression $+1$	200	98.50	98.50	+0.00	0/0	+0.00
Position progression -1	200	85.00	100.00	+15.00	30/0	+15.62
Position progression $+1$	200	88.50	100.00	+11.50	23/0	+12.33
Number arithmetic -1	200	99.50	99.50	+0.00	0/0	+0.00
Number arithmetic $+1$	200	99.00	99.00	+0.00	0/0	+0.00
Position arithmetic -1	200	100.00	100.00	+0.00	0/0	+0.00
Position arithmetic $+1$	200	99.50	99.50	+0.00	0/0	+0.00
Number Distribute Three	200	100.00	100.00	+0.00	0/0	+0.00
Position Distribute Three	200	87.00	100.00	+13.00	26/0	+13.69
Number/Position Constant	200	82.50	100.00	+17.50	35/0	+17.62
Completion count ≤ 4	916	96.62	99.78	+3.17	29/0	+3.30
Completion count > 4	1284	92.99	99.61	+6.62	85/0	+6.88
Excluding both D3 strata	1800	94.72	99.61	+4.89	88/0	+5.06
Excluding Position D3	2000	95.25	99.65	+4.40	88/0	+4.56

Scope of verification. Numerical model outputs preceded answer-label analysis. The independently authored saved-output auditor recomputes outcomes, pair counts, selected comparison formulae, allocation and primary inference from frozen records. It validates the saved computation, not an independent image-to-answer replication or a new frontend. No claim of frequency necessity or superiority over joint matching follows from these descriptive controls alone.

Table 21: Study III: PrAE low matched comparison. Base is matched normalized JS; H is expanded H with stable actual-mask inference for PrAE and averaged H for NVSA. Changes are in percentage points; C/H denotes corrections/harms. Strata and subsets are descriptive, with no separate inferential claim. An empty completion-count subset has $n = 0$.

Stratum / subset	n	Base (%)	H (%)	Δ	C/H	Tie Δ
Number progression -1	100	100.00	100.00	+0.00	0/0	+0.00
Number progression $+1$	100	100.00	100.00	+0.00	0/0	+0.00
Position progression -1	100	92.00	100.00	+8.00	8/0	+8.25
Position progression $+1$	100	94.00	100.00	+6.00	6/0	+6.00
Number arithmetic -1	100	98.00	98.00	+0.00	0/0	+0.00
Number arithmetic $+1$	100	95.00	95.00	+0.00	0/0	+0.00
Position arithmetic -1	100	100.00	100.00	+0.00	0/0	+0.00
Position arithmetic $+1$	100	99.00	99.00	+0.00	0/0	+0.00
Number Distribute Three	100	100.00	100.00	+0.00	0/0	+0.00
Position Distribute Three	100	91.00	100.00	+9.00	9/0	+9.00
Number/Position Constant	100	88.00	98.00	+10.00	10/0	+10.50
Completion count ≤ 4	1100	96.09	99.09	+3.00	33/0	+3.07
Completion count > 4	0	—	—	—	—	—
Excluding both D3 strata	900	96.22	98.89	+2.67	24/0	+2.75
Excluding Position D3	1000	96.60	99.00	+2.40	24/0	+2.47

Table 22: Study III: NVSA low guardrail. Base is normalized native; H is expanded H with stable actual-mask inference for PrAE and averaged H for NVSA. Changes are in percentage points; C/H denotes corrections/harms. Strata and subsets are descriptive, with no separate inferential claim. An empty completion-count subset has $n = 0$.

Stratum / subset	n	Base (%)	H (%)	Δ	C/H	Tie Δ
Number progression -1	100	100.00	100.00	+0.00	0/0	+0.00
Number progression $+1$	100	99.00	99.00	+0.00	0/0	+0.00
Position progression -1	100	93.00	99.00	+6.00	6/0	+6.25
Position progression $+1$	100	93.00	100.00	+7.00	7/0	+7.00
Number arithmetic -1	100	98.00	98.00	+0.00	0/0	+0.00
Number arithmetic $+1$	100	100.00	100.00	+0.00	0/0	+0.00
Position arithmetic -1	100	99.00	98.00	-1.00	0/1	-1.00
Position arithmetic $+1$	100	97.00	97.00	+0.00	0/0	+0.00
Number Distribute Three	100	97.00	97.00	+0.00	0/0	+0.00
Position Distribute Three	100	92.00	100.00	+8.00	8/0	+8.00
Number/Position Constant	100	87.00	99.00	+12.00	12/0	+12.50
Completion count ≤ 4	1100	95.91	98.82	+2.91	33/1	+2.98
Completion count > 4	0	—	—	—	—	—
Excluding both D3 strata	900	96.22	98.89	+2.67	25/1	+2.75
Excluding Position D3	1000	96.30	98.70	+2.40	25/1	+2.47

Table 23: Study III: NVSA high primary. Base is normalized native; H is expanded H with stable actual-mask inference for PrAE and averaged H for NVSA. Changes are in percentage points; C/H denotes corrections/harms. Strata and subsets are descriptive, with no separate inferential claim. An empty completion-count subset has $n = 0$.

Stratum / subset	n	Base (%)	H (%)	Δ	C/H	Tie Δ
Number progression -1	200	100.00	100.00	+0.00	0/0	+0.00
Number progression $+1$	200	97.50	97.50	+0.00	0/0	+0.00
Position progression -1	200	83.50	100.00	+16.50	33/0	+17.38
Position progression $+1$	200	87.50	100.00	+12.50	25/0	+13.08
Number arithmetic -1	200	99.50	99.50	+0.00	0/0	+0.00
Number arithmetic $+1$	200	99.00	99.00	+0.00	0/0	+0.00
Position arithmetic -1	200	99.00	99.00	+0.00	0/0	+0.00
Position arithmetic $+1$	200	99.50	99.50	+0.00	0/0	+0.00
Number Distribute Three	200	100.00	100.00	+0.00	0/0	+0.00
Position Distribute Three	200	83.50	100.00	+16.50	33/0	+17.19
Number/Position Constant	200	82.50	100.00	+17.50	35/0	+17.62
Completion count ≤ 4	916	96.07	99.45	+3.38	31/0	+3.52
Completion count > 4	1284	92.13	99.53	+7.40	95/0	+7.66
Excluding both D3 strata	1800	94.22	99.39	+5.17	93/0	+5.34
Excluding Position D3	2000	94.80	99.45	+4.65	93/0	+4.81

K.1 Complete saved-event mechanism census

After the four-partition barrier and saved-output audit, the prospectively specified census retained 474 overlapping contrast events on 184 unique questions. Table 24 reports every family, including harms and possible wrong-to-wrong disagreements. These are descriptive event counts, not 474 independent observations. Interpretation uses only the saved census records, with no new inference, scoring, XML parsing or bootstrap calculation.

Joint ideal witness and reference discipline. For the candidates actually selected by each contrast, define W to require jointly: equality of all five ideal-native components (Position, Number, Type, Size, Color); an H difference in at least one eligible attribute; and equality between the correct-completion H and the actual reference H in every eligible attribute. Ideal appearance consensus is its unanimous value or bottom. Eligibility must be nonempty. Panel indices are zero-based: PrAE and matched NVSA comparisons use panel 7; H_{avg} uses the exact average of panels 6 and 7. The correct-completion count is the count in panel $8 + y$, not the count of reference panel 7. Table 25 counts the conjunction directly, rather than inferring it from marginal flag totals.

Every matched H/P disagreement. Both backends disagree only on high-range IDs 141002718 and 141002763; all four backend-events are corrections. There are no H/P harms, wrong-to-wrong events, or further disagreement cases. Only Type is eligible, with true Constant rule. All five ideal-native components collide, including bottom Type consensus; Type supports coincide while frequencies differ, and correct H equals panel 7. Using zero-based Type values, ID 141002718 has completion/reference count $7/7$ and support $(0, 1, 2, 3, 4)$: correct/reference H is $(2, 1, 1, 2, 1)/7$, versus $(1, 1, 1, 1, 3)/7$ for the P-selected candidate. H selects correct answer 3 and P selects 5. ID 141002763 has count $5/5$ and support $(0, 3, 4)$: the corresponding histograms are $(3, 1, 1)/5$ and $(1, 3, 1)/5$, with H selecting correct 4 and P selecting 2.

The third same-support/different-H PrAE primary correction, ID 141001193, has no H/P answer disagreement. Thus a same-support primary witness alone is not evidence that frequency was necessary for its answer.

Every recorded exception: averaged NVSA. Low ID 141000481 is a correction from normalized answer 5 to correct answer 7, but its candidate ideal H/P profiles are identical. Eligible Type H is $\{1 : 1/2, 3 : 1/2\}$ for both candidates and both references. Full native collision and correct/reference equality hold, while eligible H distinction fails. The stored Type term changes the margin for 7 over

Contrast	Split	N	Gate	Corr.	Harm	W/W	Kept
PrAE H_{exp}/N	Low	1100	239	33	0	0	33
	High	2200	504	114	0	0	114
NVSA H_{avg}/N	Low	1100	240	33	1	0	34
	High	2200	505	126	0	0	126
NVSA H_7/C_7	Low	1100	240	32	0	0	32
	High	2200	505	127	0	0	127
PrAE $H_{\text{exp}}/P_{\text{exp}}$	Low	1100	239	0	0	0	0
	High	2200	504	2	0	0	2
NVSA H_7/P_7	Low	1100	240	0	0	0	0
	High	2200	505	2	0	0	2
PrAE $H_{\text{exp}}/H_{\text{old,orig}}$	Low	1100	239	52	4	0	4

Table 24: Complete predeclared family accounting. Corrections/harms are directed from the right comparator to H; W/W denotes wrong-to-wrong answer changes. Gate counts questions with any active eligible attribute. PrAE primary and H/P use stable-actual-mask inference. N in a contrast denotes its matched normalized comparator; the N column is the question denominator. The final row selects harms only; its 52 corrections were not given mechanism annotations. No high-range guardrail-harm family was specified.

Correction family	Split	Corrections	W	Support differs	Same support, H differs
PrAE H_{exp}/N	Low	33	33	33	0
	High	114	114	111	3
NVSA H_{avg}/N	Low	33	32	32	0
	High	126	126	124	2
NVSA H_7/C_7	Low	32	32	32	0
	High	127	127	125	2
PrAE $H_{\text{exp}}/P_{\text{exp}}$	High	2	2	0	2
NVSA H_7/P_7	High	2	2	0	2

Table 25: Joint witnesses among selected corrections. Support categories refer to eligible attributes of the selected pair. Support differences can coexist with frequency differences. Both low H/P families are empty. The sole correction exception is NVSA ID 141000481. The one averaged-NVSA harm fails W; all four PrAE guardrail harms have inactive gates and no positive eligible witness.

5 by $+0.0001530713$, from -0.0000588912 to $+0.0000941801$. Last consensus, last H, last P and deletion already select 7. This is not an ideal inventory-distinction witness.

Low ID 141000206 is the only averaged-NVSA harm: normalized native selects correct 2, while averaged H selects 0. Position differs, so the pair does not fully collide; its ideal H/P profiles coincide and fail to match the actual averaged Color reference. Correct-completion count is 2; panel 6 and panel 7 counts are 3 and 2. Both candidates have Color H $\{4 : 1/2, 6 : 1/2\}$, while the average reference is $\{2 : 1/6, 3 : 1/4, 5 : 1/6, 8 : 1/4, 9 : 1/6\}$. The true Position Arithmetic(-1) rule also fails the metadata gate's structural sufficient condition. Thus all four recorded exception categories apply: no full collision, no eligible H distinction, no correct/reference equality, and unjustified metadata gate. The Color term alone changes the correct-over-wrong margin by -0.0000393982, from $+0.0000127886$ to -0.0000266096. Last consensus, last H and last P all retain correct 2; last H's top margin is $+0.00190107$. The matched-last contrast does not share this harm. This localizes a saved score change without isolating reference averaging as an exclusive causal explanation.

Every recorded exception: four PrAE guardrail harms. All four losses against original old-gate H first occur from original32 to stable-legacy inference. Native, old-H and expanded-H arms share each loss. Stable-actual-mask then preserves stable-legacy actions and answers. Old and expanded gates are inactive throughout, and the active Number arithmetic rule is unchanged. The inactive Position action also changes, but is unselected in both branches. Table 27 gives all IDs and appearance changes; each carries the recorded exception "inactive gate, no positive eligible witness."

Question ID	Backend	P margin	H margin	Type increment
141002718	PrAE	-0.00370612	$+0.05792776$	$+0.06163388$
	NVSA	-0.00368425	$+0.05794964$	$+0.06163388$
141002763	PrAE	-0.02085958	$+0.08371436$	$+0.10457394$
	NVSA	-0.02080789	$+0.08376604$	$+0.10457394$

Table 26: All H/P disagreements. Margins are recorded $\mathrm{score(correct)} - \mathrm{score(Pselected)}$. Structure, Size and Color contributions are unchanged; the displayed increment is entirely in Type. No soft vectors or ideal-only scores were reconstructed.

Question ID	n_q/n_7	Answers	Number rule	Appearance changes: original to stable
141000277	$4/1$	$3 \to 0 \to 0$	Arithmetic $+$	Color: Constant $\to$ Progression -1
141000600	$4/1$	$2 \to 5 \to 5$	Arithmetic $+$	Size: Constant $\to$ Arithmetic $+$; Color: Constant $\to$ Arithmetic $-$
141000693	$2/1$	$1 \to 2 \to 2$	Arithmetic $-$	Type: Constant $\to$ Progression $+1$; Color: Constant $\to$ Arithmetic $-$
141000871	$4/2$	$5 \to 6 \to 6$	Arithmetic $+$	Type: Constant $\to$ Progression -1; Color: Constant $\to$ Arithmetic $+$

Table 27: All low PrAE guardrail harms. Answers follow original32, stable-legacy, stable-actual-mask; only the first answer is correct. n_q is correct-completion count and n_7 is reference-panel count. Gates are inactive in all branches.

For the pair consisting of the correct answer and the final wrong answer, recorded original-old-H to final-expanded-H margins are: 141000277, $+0.69298765 \to -0.00005814$ (Color increment -0.69304579); 141000600, $+0.69302958 \to -0.00005167$ (Size -0.693081257, Color $+1.76 \times 10^{-9}$); 141000693, $+0.69269144 \to -0.00020408$ (Type -0.69314580, Color $+0.00025029$); and 141000871, $+2.07925716 \to -0.00131475$ (Type -1.38611637, Color -0.69445554). Structure contributions are unchanged. These are pairwise stored margins, not necessarily top-two margins. They localize the losses to the inference transition and appearance costs, with no eligible inventory replacement active; they do not identify a causal repair.

Scope. All selected events and exceptions are retained. The census does not enumerate every possible ideal collision in all 3300 questions, establish population frequency necessity, or provide exclusive causal attribution. Shared questions and frontend also preclude treating backend-events as independent replications. Neither missing PrAE soft vectors nor native executor internals are rederived. The saved census reports zero new questions, model calls, scorer calls and bootstrap calls; its pass status certifies the bounded saved-record census, not these stronger scientific claims.

L Published visual checkpoints: transfer protocol

A fixed external input set. The public-weight experiment uses all three ObjectCNN starter states distributed with ALANS (Zhang et al., 2022), from repository commit `d853775c31acb9c8f1a25b69fdbc3942d2fea2b1`. The assets are named `sys.pth`, `prod.pth` and `loc.pth`. They are externally produced weights, not three new random-seed replications. Their checkpoint payloads do not document the training count ranges. ALANS and PrAE also use related four-branch CNN designs. The experiment changes the visual weights, while retaining the known nine-slot geometry and matched reasoning interfaces.

All 1,100 low and 2,200 high Study III questions are reused in their original order. An image-only package contains IDs and the sixteen original PNG byte strings, verified by ordered panel hashes. The questions were already opened: the prospective freeze of weights, mapping, arms and analysis defines a transfer test, not another fresh-question confirmation.

Author preprocessing and fixed semantic mapping. Each grayscale 160×160 panel is split into nine row-major crops at $(0, 53, 106, 160)$ and resized by the author's Pillow operation to 32×32, retaining the 0–255 pixel scale. ObjectCNN runs in evaluation mode with 1,024-crop batches. Synthetic checks cover cropping, batching, repetition and candidate permutation for all three states; empirical parameter and BatchNorm hashes remain unchanged.

The saved raw heads map existence class one to occupancy, Type to shared bins 1–5, Size to 0–5 and Color to 0–9; other bins are zero and the unused Angle field is a point mass. This semantic mapping has no target-fitted permutation or calibration. The shared adapter clips float32 occupancies to $[10^{-10}, 1 - 10^{-7}]$, floors categorical entries at 10^{-10} and normalizes. Raw-to-shared differences and clamp counts are retained. The experiment uses published perception with PrAE/NVSA reasoning; it is not a full native ALANS evaluation.

All controls and a single question-level estimand. Each asset is evaluated with all 21 PrAE arms and all seven NVSA arms. PrAE's high primary is stable actual-mask expanded H minus matched normalized JS; NVSA's is averaged H minus normalized native execution. Their equal-weight fixed-asset means average correctness differences over checkpoints *within each question*. The bootstrap resamples questions jointly across assets and arms, within the eleven fixed strata, using 20,000 draws. High macro intervals are two-sided 97.5%, allocating nominal 0.025 to each of the two primary backend contrasts. Positive-direction interpretation requires a strictly positive interval lower bound, at least twenty nonzero question-mean differences, within-stratum variation and no fractional-tie sign reversal. A point gain of at least one percentage point and an interval lower bound above one point are reported as separate criteria.

Per-asset intervals and low-range effects are descriptive. The low PrAE matched comparison and its final-versus-original32-old-H deployment comparison remain separate; no new low-range noninferiority claim is made. Frequency effects use reference-matched H/P arms: expanded H/P under actual-mask PrAE and last-panel H/P under NVSA. Averaged H versus last-panel P cannot isolate frequency. All-arm accuracies, paired corrections/harms, candidate changes, fractional tie credit, strata and gate coverage are retained. If an asset partition fails, its complete partition is recorded as failed and the corresponding three-asset mean remains inconclusive; no successful-subset replacement is allowed.

Execution and audit boundary. Image inference and both reasoning backends execute from a copied, hash-bound source/input bundle. The numerical backend primitives are unchanged; four documented module-level path substitutions in two copied files make the sources relocatable. Backend actions and scores are recomputed from the new public-model posteriors. A Python file-access guard rejects historical project inputs outside the bundle and current run; it is an audit mechanism, not an operating-system sandbox.

Two host-memory-reserve stops preceded the final serial continuation; both failures and partial outputs are preserved. Prospective resource amendments reduce concurrency to one and the host reserve from four to two GiB without changing scientific settings. Exact-byte reuse preserves six completed perception and eleven backend stages; seven remaining stages complete under the original aggregate four-hour budget, which includes 3,166.047 seconds from the predecessors. All twelve score partitions are sealed before answer-dependent analysis. Reused receipts identify their original computation.

A separately authored saved-output audit reconstructs mappings, winners, ties, paired counts, strata and fixed-asset contrasts, and checks percentiles against saved bootstrap draws. It does not rerun inference or generate independent draws. The hash-verified local source/input archive includes images, all public weights, licenses, environment requirements and reconstruction commands. Copied execution is tested in the recorded installed environment; independent installation is outside this check.

Portable reconstruction and saved evidence. The separately assembled portable derivative supplies a relative-path protocol and a single `release_run.py` entry point. Its 70 runtime files retain their original bytes; the analysis changes path handling and evaluation-sidecar loading while retaining all numerical definitions. Both sidecars are checked against the original Parquet labels and reference manifests after the completed audit. Artificial tests cover incomplete fixed-asset aggregation and ties. A two-root synthetic replay uses eight questions per range and the `sys` visual state: all four frontend and twelve backend worker invocations finish, and 574 arrays match byte for byte. The second root contains spaces and non-ASCII characters. This validates those worker commands in the existing installation, not the complete production launcher/analysis flow, all three forward passes in a fresh installation, or the full 3,300-question recomputation.

Table 28: **All PrAE comparisons under published visual weights.** Accuracy (%) on the same 1,100 low and 2,200 high questions. Every fixed comparison and checkpoint is included; no row was selected by its result.

	Low count			High count		
Comparison	sys	prod	loc	sys	prod	loc
Original float32						
Native	82.82	85.45	83.45	76.36	77.91	76.18
Normalized JS	82.82	85.45	83.45	76.36	77.91	76.00
Old-gate H	84.73	87.55	85.82	79.45	81.32	79.32
Old-gate P	84.73	87.55	85.82	79.32	80.77	79.14
Expanded H	85.18	88.18	86.09	80.68	82.64	80.27
Expanded P	85.18	88.18	86.09	80.59	82.09	79.95
Expanded deletion	79.64	82.09	80.18	74.50	75.86	73.14
Stable legacy-index						
Native	82.91	85.45	83.45	76.45	77.91	76.14
Normalized JS	82.91	85.45	83.45	76.45	77.91	75.95
Old-gate H	84.82	87.55	85.82	79.55	81.32	79.27
Old-gate P	84.82	87.55	85.82	79.41	80.77	79.09
Expanded H	85.27	88.18	86.09	80.77	82.64	80.23
Expanded P	85.27	88.18	86.09	80.68	82.09	79.91
Expanded deletion	79.73	82.09	80.18	74.59	75.86	73.09
Stable actual-mask						
Native	86.00	88.45	86.27	76.77	78.23	76.41
Normalized JS	86.00	88.45	86.27	76.77	78.23	76.23
Old-gate H	87.91	90.55	88.64	79.86	81.64	79.55
Old-gate P	87.91	90.55	88.64	79.73	81.09	79.36
Expanded H	88.64	91.64	89.36	81.23	83.41	80.82
Expanded P	88.64	91.64	89.36	81.14	82.82	80.45
Expanded deletion	81.91	83.91	82.09	74.77	76.18	73.36

Table 29: **All NVSA comparisons under published visual weights.** Accuracy (%) on the same 1,100 low and 2,200 high questions. Every fixed comparison and checkpoint is included; no row was selected by its result.

	Low count			High count		
Comparison	sys	prod	loc	sys	prod	loc
Native	80.27	84.91	82.64	74.09	77.36	72.73
Normalized native	80.27	84.91	82.73	74.14	77.36	72.64
Last-panel C	80.55	84.45	82.45	74.05	77.09	72.45
Last-panel H	82.82	87.82	85.55	78.59	81.73	76.55
Averaged H	82.55	88.09	85.36	78.55	81.77	76.50
Last-panel P	82.91	87.91	85.55	78.41	81.32	76.32
Eligible-term deletion	75.82	79.91	77.18	71.91	74.59	69.95

The complete saved-evidence archive is a different artifact. It retains the original absolute-path provenance and both failed resource attempts, with 11,652 dependency paths represented through 4,022 ZIP members, exact-byte aliases and a separate source/input companion. Every archived member is rehashed. Regenerable crops and interrupted predecessor arrays outside the complete-stage reuse records are omitted; those failures remain documented. The archived saved-score audit therefore checks retained predictions and arithmetic rather than independently reproducing pixels-to-answers. The portable reconstruction derivative does not alter the frozen original-run identity.

Minimal reconstruction index. The submitted supplement places the public-weight source/input kit under `image_to_score/`. Its 80 files retain their original bytes from the verified portable archive (SHA-256 prefix `b8d7cc19990d624b`). Extract to a short directory outside the installed Python packages. Run `release_run.py --preflight` with a new output directory and a

forbidden reference root supplied through `--forbid-root`; remove `--preflight` to execute. The included README gives the complete command and recorded environment; the supplement manifest lists every file hash.

Table 30: **Public-weight evidence and reconstruction map.** Each check has a distinct scope; none substitutes for the unperformed fresh-installation, full-cohort reconstruction.

Question	Locator and verified scope
Primary estimand	Table 4: matched high-count H gain, averaging the three fixed checkpoints within each of 2,200 questions; question-level 97.5% intervals.
Comparison definition	PrAE: actual-mask expanded H versus normalized JS. NVSA: averaged H versus normalized native execution. Low effects and matched-last H/P remain separate.
Complete outcomes	The preceding all-arm tables and saved analysis retain every arm, both ranges, per-checkpoint outcomes, harms, ties and strata.
Source/input reconstruction	Archive `release_protocol.json`, runtime manifest, visual weights, image-only inputs and checked evaluation sidecars; numerical analysis definitions preserved.
Tested relocation	Two-root artificial `sys` worker replay, four visual and twelve backend invocations, 574 equal arrays, recorded installed environment.
Saved-output verification	Separate audit/provenance archive verifies retained scores and paired summaries; it uses saved predictions and bootstrap draws.
Not yet executed	Fresh environment installation and full-cohort replay of the packaged launcher/analysis.

Cached scalar-comparison replay. The cached-replay archive covers all seven completed Study IV/public-extension partitions, with separated input, target and expected-output directories. It reconstructs all 17/7 comparison arms and 136/21 descriptive pairs, including all strata, fixed public assets and float32 bridge controls. The exported entrypoint reproduces 39,648 numerical arrays exactly; all winners and complete tie masks agree. Export plus replay takes 128.74 seconds (replay stage 63.77) on one CPU, with observed owned-process peak memory 139.6 MB. It starts from saved scenes, pooled H/P/G states, actions, predictions and baseline anchors. It does not repeat perception, scene construction, action inference, bootstrap or the new gate/census analysis.

What is supplied. The supplement also contains `cached_comparison_code/`, with the unchanged nine runtime Python files and upstream license from the cached replay, and complete descriptive summaries for all Study IV/public-extension arms and pairs. The recorded cached environment is Python 3.12 with NumPy 2.5.2. The full 328 MB cached archive is retained locally; its 3,308 numerical files are not included in the 100 MB-limited supplement, so the supplied scalar source alone is not a self-contained replay. The separate `image_to_score/` kit includes its images and weights but has the tested relocation scope stated above. LaTeX sources are under `paper_source/`. These files expose the implementations and primary image inputs; they do not constitute an executed fresh-installation, full-cohort reconstruction of every study.

M PUBLIC-WEIGHT CANDIDATE MECHANISM CENSUS

A saved-output census retains all three visual states and both ranges: 393 unique questions contribute 1622 overlapping event memberships across the predeclared families. The unit of deduplication is the question ID and range; checkpoint and backend repetitions are not new questions. For the same selected candidate pair, an aligned witness requires a nonempty eligible gate, equality of the complete ideal native summary, an eligible H difference, a metadata-compatible gate, and correct-completion H equality with the relevant reference. These are event-conditioned compatibility checks, not a claim of exclusive causation. All harms, nonwitness corrections, ties and wrong-to-wrong H/P changes remain in the saved census. The parser preserves exact rational H/P inventories and distinguishes reference-panel count (panel 7) from correct-completion count (panel $8 + y$). Saved component margins are reconstructed at tolerance 10^{-12}. Full first-index, tie-set and fractional-credit records cover all original arms; ideal collisions need not imply equal soft costs. The completed census took 56.4 seconds with no generator, model, scorer or bootstrap calls. An initial 280.312-second resource

Table 31: **Inventory corrections under published visual weights.** C/H lists all corrections/harms. The final columns show aligned joint witnesses among corrections in the same contrast. Low and high denominators are 1,100 and 2,200 questions per checkpoint; families overlap.

Comparison	Weights	Low C/H	High C/H	Low aligned/C	High aligned/C
PrAE H / norm	sys	30/1	118/20	21/30	66/118
PrAE H / norm	prod	38/3	137/23	30/38	73/137
PrAE H / norm	loc	35/1	121/20	28/35	61/121
NVSA avg H / norm	sys	27/2	112/15	17/27	71/112
NVSA avg H / norm	prod	37/2	112/15	28/37	74/112
NVSA avg H / norm	loc	29/0	100/15	26/29	68/100
NVSA last H / C	sys	28/3	118/18	16/28	72/118
NVSA last H / C	prod	40/3	122/20	31/40	75/122
NVSA last H / C	loc	34/0	105/15	29/34	68/105

Table 32: **Reference-matched H/P disagreements.** C/H/W lists corrections, harms and wrong-to-wrong changes. Frequency/C counts corrected events with an aligned same-support, different-frequency ideal witness. A zero denominator denotes no correction; it does not establish equivalence of scores or tie credit.

Comparison	Weights	Low C/H/W	High C/H/W	Low frequency/C	High frequency/C
PrAE H / P	sys	0/0/0	7/5/3	0/0	2/7
PrAE H / P	prod	0/0/0	14/1/0	0/0	1/14
PrAE H / P	loc	0/0/0	13/5/1	0/0	0/13
NVSA last H / P	sys	1/2/0	9/5/3	0/1	4/9
NVSA last H / P	prod	1/2/1	11/2/2	0/1	1/11
NVSA last H / P	loc	0/0/0	10/5/0	0/0	1/10

timeout is preserved. The completed attempt caches Windows child-process membership for 250 ms while retaining live own/known-child RSS checks, all input hashes, the unchanged external monitor and every scientific function. It does not reuse partial census outputs or change the selected cases.

Available information and soft-score effects. A separately implemented saved-record check reconstructs the same-pair flags from exact rational object profiles, verifies all 30 family selections and 330 stratum summaries against the complete score analysis, and recomposes the recorded margin changes. Among high-count primary corrections, 176/376 PrAE and 111/324 NVSA events lack the full joint witness. The failed conditions overlap: metadata compatibility fails in 157/176 and 92/111 of these nonwitness corrections, correct-to-reference equality in 154/176 and 92/111, and complete native collision in 92/176 and 67/111. Ten events in each backend have no eligible ideal H distinction. These are overlapping checks on selected events, not disjoint causes or estimates of population prevalence. Every selected event has a nonempty inferred gate.

These metadata checks are sufficient conditions; an unsupported check does not prove that an inferred action is incorrect. The joint witness also holds for three of the 63 high PrAE harms and one of the 45 NVSA harms. An ideal distinction can therefore be available while imperfect soft evidence changes the answer in the wrong direction. The information-loss result characterizes what a representation retains; it does not certify the sign of every candidate-score intervention.

Frequency cases without repeated-question inflation. The aligned H/P frequency corrections comprise nine event memberships but only five high-range questions: `141001168`, `141001811`, `141002499`, `141002763` and `141003229`. The last-but-one also appears in Study III's two-question frequency set. The other Study III question, `141002718`, appears among the public-weight audit's raw same-support/different-frequency corrected pairs but fails the full alignment conjunction there. A raw frequency contrast alone is therefore not counted as an aligned witness. One aligned `loc`/PrAE H/P event on `141003229` changes an incorrect answer to another incorrect answer; it is retained separately from the corrections. The complete saved table includes all such answer changes, not only successful cases.

Table 33: **Source and opened development evidence.** Source uses one fixed perception model on 1,000 low-count questions; opened high averages three models on another 1,000 shared questions. The seven original arms are retained. Presence and deletion are later source-only controls; dashes denote evaluations not performed. None of these columns is the fresh confirmation cohort.

Score	Source accuracy (%)	Opened high accuracy (%)
Native PrAE	93.80	93.53
Normalized JS	93.90	93.57
Attribute histogram	95.40	97.80
Unaligned slots	90.60	93.13
Spatial correspondence	95.40	97.93
Independent attribute OT	95.40	97.80
Joint object OT	95.40	97.93
Clipped-count presence	95.40	—
Eligible-term deletion	88.80	—

N Complete matching control definitions

To separate marginal inventories from object binding, let $\Gamma(w_7, w_j)$ be the set of couplings with source and candidate occupancy marginals. The local cost for attribute a is $c_{sta} = \mathrm{JS}(q_{7sa}, q_{jta})$. Independent matching and joint matching replace the eligible total by, respectively,

$$D_I(j) = \sum_{a\in\mathcal{A}} \min_{\gamma\in\Gamma(w_7,w_j)} \sum_{s,t} \gamma_{st} c_{sta}, \quad D_J(j) = \min_{\gamma\in\Gamma(w_7,w_j)} \sum_{s,t} \gamma_{st} \sum_{a\in\mathcal{A}} c_{sta}. \tag{11}$$

Independent matching can assign different partners for different attributes. Joint matching keeps the attributes of an object together. The spatial control instead uses $T(s) = (s + d) \bmod 9$, with $d \in \{0, 1, -1, 2, -2\}$ determined by the selected action, and sums $w_{7s} c_{s,T(s),a}$. An unaligned control fixes $T(s) = s$. These spatial arms use source weights only; unlike OT, they do not impose candidate occupancy marginals.

O Earlier source and opened-development controls

O.1 What the development controls established

On source validation, histogram and all three structured matching arms make identical decisions: 15 corrections and no harms relative to normalized JS. The original spatial-specificity screen required improvement over the histogram and matching controls; it failed. This motivated a separate fresh comparison of the common distribution-retention effect, without changing the outcome of that screen.

The two later source controls sharpen this interpretation. Histogram and presence both answer 954 questions correctly and agree on every decision, although their scores differ by up to 0.205. Deletion answers 888 correctly: it loses 52 previously correct normalized-JS answers and recovers one. Relative to deletion, histogram recovers 66 answers and loses none. The added comparison therefore supplies useful answer evidence on these questions. Presence remains an equally accurate simpler comparator here; these data do not establish that preserving frequency is necessary. Uniform tie credit gives 952.50 correct-answer equivalents for histogram and presence, versus 885.625 for deletion, so the source ordering also holds under the declared tie convention.

Inventory preservation extends beyond a fixed movement. The gate in the fresh experiment was inherited from a spatial-comparison design. A subsequent inspection of the pinned generator identified another inventory-preserving branch: Position Distribute Three changes occupied locations while retaining Constant-attribute values within a row. A separately specified source pilot added only this branch to the native prediction-based gate, keeping the previous nine scores unchanged. It activated 13 additional questions and increased histogram accuracy from 95.40% to 95.80%: four corrections, no harms. Expanded presence, independent OT and joint OT made the same decisions; matched deletion reached 88.60% and unaligned comparison 90.40%. All four corrections occurred

among the 91 true Position Distribute Three questions, whose accuracy increased from 58.24% to 62.64%. This exploratory result motivates separating the inventory-preservation condition from the availability of a fixed slot correspondence. It is not part of the preceding fresh confirmation.

On the opened high-count cohort, histograms reach 97.80%, compared with 93.57% for normalized JS (Table 33). Spatial and joint matching make identical decisions on every model and reach 97.93%. Their additional benefit over histograms is small. In two visual-model-0 examples, the wrong candidate has the correct marginal inventories but different object attribute pairings; joint and spatial comparison recover the correct answer. These cases establish a concrete remaining histogram ambiguity, not a spatial-specific advantage. Appendix D reports all model counts and a binding example.

The unaligned control is worse than normalized JS on source and opened development (Table 33); the fresh high-count aggregate instead improves slightly. It pairs slots without applying the selected movement, so its effect depends on the questions' selected relations. Arbitrary slot pairing is insufficient for a movement rule, but the stronger unordered controls remain the relevant comparison for spatial structure.

P Execution receipts and numerical optimization

Table 34: Study I stage wall times, in seconds, over three model processes (minimum–maximum). Native CPU jobs run at most two concurrently; these are execution receipts, not isolated latency benchmarks. Seven-arm scoring includes native replay and all OT controls.

Stage	Low: 1,100	High: 2,200
Perception	5.5–8.7	8.1–8.2
Full native inference	675.5–796.4	1286.3–1502.1
Cached seven-arm scoring	22.9–24.1	44.9–60.5

Table 34 separates perception, full native execution and cached scoring on the local RTX 5070 Laptop system. Process setup is included. The scores share native actions; rescoring time does not replace the cost of acquiring them. Study III took 3,219.8 seconds from fresh generation through both backend evaluations, paired analysis and saved-output audit. This total includes shared preparation and is not an isolated backend latency comparison. The histogram comparison uses no optimizer. OT controls solve an unregularized 9×9 transport problem with 81 variables and 18 equality constraints. Marginal masses are multiplied by 10^6 for the solve, then plans and primal objectives are returned to probability units. HiGHS primal and dual feasibility tolerances are 10^{-9}. Every active solution retains a primal plan and dual certificate; feasibility, objective gap and reduced-cost residuals are checked in probability units at 10^{-7}. Joint matching uses summed attribute JS costs.

Q Scalar-rule candidate conditioning: definition and fresh Study IV

Q.1 Two different comparison requirements

Eligible Constant rules preserve an object's value while the Position operation preserves the object list. H therefore compares possibly heterogeneous inventories. The generator's non-Constant appearance rules instead make each panel homogeneous in the affected attribute. The native progression, arithmetic and D3 transitions emit only ordinary attribute values. Study IV changes candidate comparison for these inferred actions, retaining the existing Constant gate, structural score, perception, inferred actions and transition tables. The gate uses predictions, including incorrect predictions; true rules never select evaluation cases.

For ordinary mass $r > 0$, write a proper candidate law as $(rc, 1 - r)$ and the executed prediction as $(a, 0)$. C-output compares a and c by ordinary JS. H candidate-only compares the same executed prediction to $(H, 0)$; its retained-mass control compares to $(rH, 1 - r)$. All four costs are composed with the same existing S_H score, replacing only non-Constant appearance terms. The shared-prediction four-cell analysis uses the native64 prediction in every cell. The preserved C-output arm

conditions the original float32 prediction; C-input re-execution uses float64. These paths are retained as separate numerical controls.

Q.2 Exact divergence decomposition

Define $\mathrm{JS}_w(a,c) = w\,\mathrm{KL}(a\|m_w) + (1-w)\,\mathrm{KL}(c\|m_w)$, where $m_w = wa + (1-w)c$. Then

$$\begin{aligned} \mathrm{JS}((a,0),(rc,1-r)) &= B(r) + \frac{1+r}{2}\,\mathrm{JS}_{1/(1+r)}(a,c), \\ B(r) &= \log 2 + \frac{r}{2}\log r - \frac{1+r}{2}\log(1+r). \end{aligned} \tag{12}$$

Partition the mixture into ordinary and inconsistency events. Its ordinary mass is $(1+r)/2$ and its conditional law is $(a+rc)/(1+r)$. The entropy chain rule gives the binary contribution $h((1+r)/2) - h(r)/2 = B(r)$ and the displayed weighted categorical term. Thus conditioning removes a consistency discrepancy *and* changes the categorical mixture weight; it is not merely subtracting an additive penalty. At $r = 1$ the costs coincide; as $r \downarrow 0$, the native cost tends to $\log 2$ regardless of c. Conditioning is undefined at zero.

For known occupied objects with identical heads u and count N, the ordinary mass is $r = \sum_v u_v^N$. It decreases with N for a nondegenerate head, but $c_v = u_v^N / \sum_w u_w^N$ also changes. Conditioning therefore does not remove all count dependence or establish calibration. Consistency may be useful: for two candidates with $c = a$ but different r, conditioning ties them while native JS prefers the larger mass. A wrongly inferred scalar rule can likewise discard valid evidence that a panel is heterogeneous. These are limits of the method, not failures of the decomposition.

Q.3 Why conditioning execution is redundant

Let input laws be $x = (s_1 u, 1-s_1)$ and $y = (s_2 v, 1-s_2)$ with positive ordinary masses. For unnormalized native execution E,

$$E(x,y) = \kappa E((u,0),(v,0)), \qquad \kappa = \begin{cases} s_2 & \text{Progression or D3-left,} \\ s_1 & \text{D3-right,} \\ s_1 s_2 & \text{Arithmetic.} \end{cases} \tag{13}$$

Progression and D3-left sum the first input over every state, including inconsistency; D3-right sums the second. Arithmetic uses ordinary entries from both inputs. All emit zero inconsistency mass, and normalizing a positive output cancels κ. The original bounded support and selected panels are unchanged. This statement does not apply to Constant execution or imply bitwise equality between float32 and float64 paths.

Q.4 Development, prospective freeze and complete outcomes

The source pilot evaluated all 1,000 existing validation questions. Its primary, H input re-execution, had to gain at least two answers over each of six controls, correct at least twice as many errors as it harmed, and improve tie credit. It failed against P and C. C-output and H candidate-only each corrected eight baseline errors without harms but disagreed on four answers. A separate prospective study then chose the simpler C-output intervention as its sole primary; it did not rename the failed source primary.

Before generation, Study IV froze 3,300 low-count questions, 300 in each of eleven structural strata, all panels containing one to four objects. One existing local visual0 detector and stable actual-mask PrAE inference were fixed. All 17 arms were scored before a separate process opened answers and true-rule metadata. Parquet and generation-manifest labels agreed exactly. Five historical identity checks excluded known duplicates, including all 3,819 accepted and rejected raw attempts from Study III. The index additionally retained two synthetic fixture identities as conservative exclusions; it cannot exclude undocumented runs.

The single primary is C-output minus expanded H, with equal stratum weights, 20,000 shared within-stratum bootstrap draws and a two-sided 95% linear-percentile interval. Confirmation requires a positive lower bound, at least 20 nonzero paired correctness differences, within-stratum variation, a

nondegenerate draw distribution and positive tie-credit gain. The practical point target is $+0.5$ points, reported separately. Near ties within 10^{-6} divide credit equally. No post-result sampling, changed primary or alternate significance test was used. Other contrasts are descriptive.

The primary gains $15/3300 = 0.4545$ percentage points: 25 corrections, 10 harms and two wrong-to-wrong changes. Its 95% interval is $[0.1212, 0.8182]$ points, and its 35 discordant correctness outcomes pass the non-sparsity conditions. Tie credit gains 0.5000 points, interval $[0.1515, 0.8636]$. Statistical confirmation passes; the literal 0.5-point practical target is not reached. This does not establish a practically negligible effect. No high-count or public visual-weight cohort was evaluated in Study IV.

Retaining mass while changing C to H corrects one answer and harms none. Conditioning C corrects 25 and harms 10; conditioning H relative to its retained mass version corrects 21 and harms none. H candidate-only corrects 22 anchor errors without harms. Against unit-mass C, however, it corrects 12 and harms five: the seven-answer advantage has only 17 correctness discordances and is not the frozen primary. C execution and C-output have identical winners and tie sets. The common-prediction C cell differs from preserved C scores by at most 4.44×10^{-16} and changes no winner, tie set or fractional credit. Native64 alone changes no anchor answer. Deleting non-Constant terms yields 35.88%, so removing appearance evidence does not explain the gains.

Q.5 Execution and saved-result verification

Generation through complete scoring and analysis took 606.8 seconds: one GPU perception pass took 14.6 seconds, and the single CPU reasoning pass took 300.1 seconds. All 413 batches, including the final four questions, were retained. There were no training updates. The independent saved audit passed 67,898 checks in 7.3 seconds, verifying all 17 score paths, the common-prediction four cells, all 136 pairwise summaries in twelve scopes, two-source labels and the saved-bootstrap interval endpoints. It did not repeat model inference or draw new bootstrap samples.

The first audit stopped on an overly strict bitwise-equality assertion: identical fractional-credit input columns produced saved draw values differing by at most two ULP (2.84×10^{-14} percentage points). The preserved versioned repair uses a derived float64 bound $200\gamma_{320} = 7.11 \times 10^{-12}$ points and requires equality on the exact credit-count lattice. An independent artificial integer/rational fixture verified the contract. Neither scores, bootstrap draws, interval endpoints nor scientific decision rules changed. An earlier metadata preparation check had included an unrelated concurrently edited scorer; its preserved repair reused the unchanged identity index without generating data.

The separate Study IV records contain the complete scores, state laws, predictions, costs, action choices, ties, every positive/equal/negative pair and versioned failures. The earlier public-weight reconstruction archive covers the earlier transfer study; it is not a packaged reconstruction of this extension.

Table 35: **Complete Study IV panel.** All arms share stable actual-mask inference and the same Constant H terms where applicable. Source is development, not confirmation. "Execution" replaces both context representation and candidate representation; one-sided controls separate these changes.

Comparison	Source / 1,000	Fresh / 3,300	Fresh accuracy (%)
Native score	965	3172	96.12
Normalized JS	966	3172	96.12
Original-gate H	981	3234	98.00
Original-gate P	981	3234	98.00
Expanded H (anchor)	989	3261	98.82
Expanded P	989	3261	98.82
Constant-term deletion	900	2973	90.09
H execution	996	3282	99.45
P execution	995	3281	99.42
G execution	994	3278	99.33
C execution	997	3276	99.27
Native64 execution	989	3261	98.82
H prediction only	990	3264	98.91
H candidate only	997	3283	99.48
Non-Constant deletion	341	1184	35.88
C-output (primary)	997	3276	99.27
H candidate, retained mass	—	3262	98.85

Table 36: **Candidate shape versus ordinary mass.** All four cells use exactly the same native64 executed prediction. Counts are correct answers out of 3,300. These are descriptive mechanism controls.

Candidate shape	Retain native mass r	Unit ordinary mass
Conditional consensus C	3,261	3,276
Inventory H	3,262	3,283

Table 37: **Every Study IV stratum.** Each row has 300 questions. The table retains both strata in which the primary loses accuracy. H candidate-only remains a secondary comparison.

Structural rule	Expanded H	C-output	H candidate-only
Number progression -1	300	300	300
Number progression $+1$	298	300	300
Position progression -1	298	298	298
Position progression $+1$	298	300	298
Number arithmetic -1	295	298	298
Number arithmetic $+1$	291	298	299
Position arithmetic -1	300	300	300
Position arithmetic $+1$	288	289	293
Number D3	294	297	298
Position D3	300	299	300
Number/Position Constant	299	297	299

R CANDIDATE COMPARISON WITH PUBLISHED VISUAL CHECKPOINTS

R.1 A FIXED, DESCRIPTIVE TRANSFER PROTOCOL

The extension reuses all opened Study III questions: 1,100 low-count and 2,200 high-count questions under each of the same three published ObjectCNN weight states. The seven-arm panel, numerical bridge, inputs and analysis were frozen before rescoring. This is a test under weaker fixed perception, not another fresh cohort or a different visual architecture. Public training count ranges remain undocumented.

Each arm starts from expanded Constant H with stable actual-mask PrAE actions. The gate is the inferred non-Constant appearance action, including incorrect actions. All other terms and Constant comparisons stay fixed. Native scenes are reconstructed from saved object predictions with the archived adapter and scene engine. Three small appearance transitions then supply one common float64 prediction. Native64 compares the full native candidate, C1 its conditional ordinary shape, H1 the inventory, Hr that inventory with native ordinary mass and bottom, and P1 the matched clipped-presence candidate. Del sets precisely the same non-Constant costs to zero. H/P candidates share execution and the Constant H anchor; they do not change the context representation.

The preserved C-output bridge instead conditions the original float32 executed prediction, matching Study IV's primary. It is a numerical diagnostic, not an additional selectable arm. Every score partition completes before the separate analysis reads either target source. Analysis averages the three fixed assets within each question. The 9,900 asset–question events are correlated evaluations of 3,300 questions. All seven arms, 21 pairs, eleven strata, per-asset outcomes and tie-credit changes are retained in the saved analysis and full paired report. Tables below show all arms and strata. No bootstrap, significance test, model selection or fresh-confirmation claim is made.

R.2 COMPLETE OUTCOMES AND INTERPRETATION

C1 improves every checkpoint in both ranges. Its average gains are 1.27 points at low counts and 1.98 at high counts. High-count changes include 236 corrections and 105 harms; low-count changes include 74 and 32. These counts are overlapping asset events. Native64 changes no answer or tie set. The C-output bridge has identical winners, tie sets and credits to C1 for all 9,900 evaluations, although its scores differ by up to 3.73×10^{-8}.

At high counts, replacing C by H at native mass adds 0.03 points; setting unit mass adds 1.98 points with C and 2.05 with H. H1 exceeds C1 by only 0.09 points (98 corrections, 92 harms), and is below C1 for the prod checkpoint. At low counts H1 is 0.12 points below C1. H1 exceeds P1 by 0.27 high-count points and 0.03 low-count points; this descriptive difference alone does not identify a frequency witness. P minus C at high counts is -0.18 points by first-index accuracy but $+0.07$ by tie credit; these small rankings depend on the metric. Deletion is substantially worse in both ranges. The evidence supports useful ordinary-value comparison, with gains driven mainly by changing candidate mass, rather than a universal preference for one candidate shape.

The aggregate gain does not imply stratum-wise dominance. C1 loses 0.67 points in each of high-count Number progression -1 and Position arithmetic -1, and 1.00 point in low-count Position arithmetic $+1$. These outcomes are retained in Table 39. The extension therefore remains an empirical comparison choice with measurable harms.

R.3 EXECUTION AND VERIFICATION

The complete cached comparison took 162.3 seconds under the fixed 900-second cap: scoring 126.9, analysis 16.3 and independent saved audit 18.8 seconds, including process supervision. Peak observed owned-process RSS was 665 MB. No image generation, CNN forward, rule abduction or training was repeated. All 1,239 batches and six partitions completed. The separately implemented audit passed 248,906 checks, including native component reconstruction, exact anchor scores, independent transition arithmetic, all seven score paths, numerical bridge, two-source targets and every saved statistical summary. Its scope begins at the saved float32 scenes; it does not independently regenerate the 511-mask scene engine or perception. Protocol and result identities are retained with the local evidence. The previous portable archive does not yet reconstruct this new cached extension.

Table 38: **All candidate comparisons with published weights.** Accuracy in percent. Each range mean averages three fixed checkpoints within the same question. The old anchor retains the previously evaluated Constant H comparison.

	Low count, 1,100 questions				High count, 2,200 questions			
Comparison	sys	prod	loc	Mean	sys	prod	loc	Mean
Constant H anchor	88.64	91.64	89.36	89.88	81.23	83.41	80.82	81.82
Native64	88.64	91.64	89.36	89.88	81.23	83.41	80.82	81.82
C, unit mass	89.18	92.45	91.82	91.15	81.73	85.23	84.45	83.80
H, unit mass	88.91	92.64	91.55	91.03	82.23	84.86	84.59	83.89
H, retained mass	88.64	92.00	90.00	90.21	81.09	83.55	80.91	81.85
P, unit mass	88.73	92.64	91.64	91.00	81.95	84.59	84.32	83.62
Non-Constant deletion	38.82	37.73	35.55	37.36	39.45	38.23	37.09	38.26

Table 39: **All structural strata in both ranges.** Each cell is the mean accuracy across three fixed checkpoints. Low/high strata have 100/200 questions. Every arm uses inferred appearance actions, not these true structural labels.

Structural rule	Anchor	N64	C1	H1	Hr	P1	Del
Low count							
Number progression -1	84.33	84.33	85.67	85.00	84.67	85.00	32.33
Number progression $+1$	85.33	85.33	89.00	89.33	85.00	88.67	30.33
Position progression -1	92.33	92.33	93.33	93.00	93.00	93.00	42.33
Position progression $+1$	90.67	90.67	91.00	91.67	91.67	92.00	42.67
Number arithmetic -1	91.00	91.00	92.67	91.33	91.33	91.33	34.33
Number arithmetic $+1$	90.00	90.00	90.67	91.67	89.33	91.00	27.33
Position arithmetic -1	93.00	93.00	94.67	93.67	93.33	93.67	42.00
Position arithmetic $+1$	90.67	90.67	89.67	91.33	92.33	92.33	38.33
Number D3	91.67	91.67	93.33	92.00	92.00	92.00	30.33
Position D3	89.67	89.67	92.00	92.00	89.33	91.67	46.00
Number/Position Constant	90.00	90.00	90.67	90.33	90.33	90.33	45.00
High count							
Number progression -1	80.67	80.67	80.00	81.67	80.33	81.33	32.17
Number progression $+1$	80.83	80.83	84.00	84.00	81.50	83.83	31.17
Position progression -1	85.33	85.33	88.00	87.17	85.00	86.67	47.50
Position progression $+1$	85.17	85.17	86.17	87.17	85.33	87.00	46.00
Number arithmetic -1	82.17	82.17	82.83	82.17	81.00	82.17	34.33
Number arithmetic $+1$	79.83	79.83	87.17	86.33	81.17	84.33	32.33
Position arithmetic -1	84.17	84.17	83.50	83.67	84.00	84.17	42.17
Position arithmetic $+1$	82.17	82.17	85.17	85.67	82.33	85.33	35.83
Number D3	80.00	80.00	80.83	80.83	80.17	81.17	32.67
Position D3	81.00	81.00	83.17	83.33	80.50	82.83	42.83
Number/Position Constant	78.67	78.67	81.00	80.83	79.00	81.00	43.83

S STRUCTURAL ELIGIBILITY VERSUS APPLYING EVERY INFERRED CONSTANT

We freeze eight cells crossing H/P comparison, the existing structural gate/all inferred Constant appearance rules, and original/C-output scalar comparison. All perception, actions, structural costs and unaffected appearance costs are unchanged. This is a descriptive reuse of the opened Study III questions with all three public visual weights; no model, threshold or gate is selected from these results.

For each newly admitted Constant term, we add the saved native cost minus its saved H/P comparison cost to the matched gated score. The same C-output difference is used in both gate settings. Saved H, P and C-output anchors are preserved; an empty new gate gives identical scores. All six partitions and all 28 pairs are retained in the result archive. No bootstrap or significance test is applied.

Table 40: Gate ablation. Accuracy is the mean across three fixed visual weights on the same 1,100 low/2,200 high questions. All Constant refers to inferred appearance actions, not annotated rules.

Representation	Constant gate	Scalar comparison	Low	High
H	Structural	Original	89.88	81.82
H	Structural	Conditioned	91.15	83.80
H	All Constant	Original	90.36	82.35
H	All Constant	Conditioned	91.88	84.39
P	Structural	Original	89.88	81.47
P	Structural	Conditioned	91.18	83.61
P	All Constant	Original	90.45	82.08
P	All Constant	Conditioned	91.82	84.26

Table 41: All-Constant minus structural-gated comparison. C/H counts are corrections/harms across correlated checkpoint–question events. Tie differences use equal credit over candidates within 10^{-6} of the maximum.

Representation	Scalar	Low Δ	Low C/H	High Δ	High C/H
H	Original	+0.48	37/21	+0.53	178/143
H	Conditioned	+0.73	46/22	+0.59	184/145
P	Original	+0.58	40/21	+0.61	181/141
P	Conditioned	+0.64	43/22	+0.65	192/149

The broader gate improves both aggregate ranges for all four prespecified contrasts. With H and C-output it changes low/high means from 91.15/83.80% to 91.88/84.39%; the respective gains comprise 46/184 corrections and 22/145 harms. Its tie-credit gains are $+0.58/+0.65$ points. Thus the structural gate is not empirically necessary or accuracy-optimal here. It supplies a sufficient object-preservation condition. Outside it, homogeneous or coincident inventories can in principle remain comparable; these score summaries do not identify that as the cause of the gain. Effects also vary across assets and strata: high-count sys is $-0.14/0.00$ points for original/conditioned H, and Number Progression $+1$ is negative in all four gate contrasts (-1.67 to -3.17 points). The original gate is retained for the frozen primary studies; this follow-up does not replace their method after inspecting outcomes.

The audit reconstructs all eight score formulas, all accompanying gate arrays, literal winners, full tie masks, exact 840-unit tie credits, and every asset/stratum summary. It passes 155,900 assertions. The complete path takes 57.58 seconds using cached inputs on one CPU; this is not end-to-end inference time.

T SCALAR CONDITIONING: COMPLETE CORRECTION AND HARM CENSUS

We enumerate every changed first-index winner or changed full 10^{-6} tie mask for the frozen C-output comparison in Study IV and the public-weight follow-up. Selection uses saved scores alone; symbolic annotations are read afterwards. Across 13,200 checkpoint–question evaluations (6,600 distinct questions), the census contains 578 events from 467 distinct questions. Every selected event is

retained, including wrong-to-wrong and tie-only changes. Metadata and original Parquet targets agree, and totals reproduce all seven original partition reports.

Table 42: Complete scalar-change accounting. Public rows aggregate correlated events across three weights on shared questions. Tie-only means the first winner is unchanged but the full tie mask differs.

Cohort	Corrected	Harmed	Wrong–wrong	Tie-only	Total
Fresh low	25	10	2	5	42
Public low	74	32	6	15	127
Public high	236	105	12	56	409

Rule recovery and uncertainty are distinct. For every active Type, Size or Color comparison we record the inferred rule's family and parameter, the corresponding generator metadata, and compatibility with all eight ideal context panels. Distribute Three metadata omit orientation: family matching is recorded separately from left/right compatibility reconstructed from context. Accidental context compatibility is distinct from metadata agreement.

Public high-count corrections contain 505 active attribute-event pairs: 495 match the available metadata fields and 497 satisfy the ideal context. Harms contain 205 pairs, with 198 metadata matches and 200 compatible contexts. Thus incorrect inferred rule identities alone cannot explain the harms. These are attribute-event counts among changed answers, not whole-cohort rule-recovery rates.

Every target, old winner and new winner is annotated homogeneous for the active attributes in both groups of public high-count corrections and harms. Their soft consistency masses nevertheless differ. For Color-active corrections, target ordinary mass has median 0.742 (148 events); for Color-active harms it is 0.922 for the target and 0.654 for the new wrong winner (56 events). Here the inconsistency mass is a model belief, not a record of actual heterogeneity in these candidate roles. Conditioning can remove a penalty on a visually uncertain correct value, but can also promote an uncertain wrong value. These medians describe the selected event populations; they do not estimate all-question prevalence or prove a unique cause.

Where the ranking changes. Each saved event includes all eight scores, all three appearance costs, all 16 panels' ordinary masses and exact old/new candidate margins. For candidates u, v, the measured margin change is

$$\begin{aligned}&[S_{C|\mathcal{V}}(u) - S_{C|\mathcal{V}}(v)] - [S_H(u) - S_H(v)]\\ &\quad= \sum_{a:\hat{r}_a \neq \text{Constant}} \{[D_{Ca}(u) - D_{Ca}(v)] - [D_{C|\mathcal{V},a}(u) - D_{C|\mathcal{V},a}(v)]\}.\end{aligned}$$

We verify this identity for new-versus-old, target-versus-old and target-versus-new pairs. The binary term $B(r)$ and native categorical remainder are recorded only for active non-Constant attributes, where the executed prediction has zero inconsistency mass. Conditioning changes both $B(r)$ and the categorical mixture weights; we do not identify the full score change with $B(r)$ alone.

Public high-count harms have median target-versus-best-wrong margin $+0.00631$ before conditioning and -0.0000627 afterwards; corrections move from -0.0231 to $+0.00250$. The complete census retains ties and large-margin failures as well. Fresh-local corrections differ: only 33/54 active attribute-event pairs match the recorded rules, compared with 19/23 among harms. Rule and uncertainty explanations should therefore not be transferred between these perception settings without checking their annotations.

In the correction, Color Distribute Three predicts the target's true value, but the target has lower soft ordinary mass than the old wrong winner. Conditioning reverses the literal ranking, largely through Color; the resulting margin is below 10^{-6}, so this is a tolerance tie with fractional credit improving from zero to one half, not decisive separation. In the harm, the Color Progression -2 rule is also correct, yet conditioning promotes the wrong value with lower ordinary mass. The target loses its 0.05274 margin and fractional credit falls from one to zero. Type and Size contributions oppose that flip. The examples illustrate the two directions of the same score intervention; they are not selected for large effects.

Table 43: Deterministic examples: the lowest question ID in each public-high event class, breaking asset ties in sys/prod/loc order. Let u be the new winner and v the old winner; margins are $S(u)-S(v)$. Both use sys: correction ID 141001108 and harm ID 141001113. Candidate indices are zero-based. All selected true appearance attributes are homogeneous.

Event	$v \to u$	Color r_u/r_v	Color Δ	Old margin	New margin
Correction	$1 \to 3$	$.998768/.999650$	$+3.059\,10^{-4}$	$-3.057\,10^{-4}$	$+3.946\,10^{-9}$
Harm	$1 \to 7$	$.852066/.999909$	$+5.420\,10^{-2}$	$-5.274\,10^{-2}$	$+5.849\,10^{-6}$

A separately authored standard-library audit reconstructs all event labels, ties, exact credit, margins, $B(r)$ and every reported semantic/continuous summary from the saved event records, and checks their completeness against the original paired reports. It passes 260,990 assertions with maximum score/margin accounting error 2.22×10^{-16}. Raw XML is independently checked for the two fixed examples; the remaining symbolic extraction uses the separately reviewed source. This audit does not rerun perception, scene construction or JS scoring. The complete census takes 133.81 seconds under its resource controller; the saved-event audit takes 1.12 seconds.