

# WHAT DOES SPATIAL ACCURACY CERTIFY? MATCHED AND CROSSED CANDIDATE AUDITS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Spatial multiple-choice accuracy can conceal candidate shortcuts and unstable answers under controlled layout changes. We test what such scores certify using question-blind pixel rules, crossed scene–instruction cases, matched foil translations and repeated responses across six model configurations and 47,136 formal requests. A fixed question-blind rule solves 245/326 released STARE questions, though candidate-only prompting performs poorly on a matched subset. Our balanced construction gives exact 1/4 accuracy ceilings for specified input-omission rules. Yet Qwen2.5-VL-32B (Qwen32) original preprocessed inputs admit a perfect query-plus-centroid rule; direct-grid rendering restores seven declared source and native bounds. On one aligned 78-cell shape, Qwen32 rejects both omission nulls, but 54.27% of cross-layout response pairs disagree, versus 2.05% of same-input pairs. An exact categorical permutation test rejects response-distribution equality; the mean half-squared categorical distance is 0.5222 (95% interval [0.4586, 0.5859]). Separately, the simultaneous 95% interval for Qwen32’s disk-pair mean-accuracy change is  $[-3.18, 3.48]$  points, within the prespecified  $\pm 5$ -point margin despite 397 correctness reversals among 2,048 pairs. These results separate evidence for required inputs, mean tolerance and response stability without identifying an internal spatial algorithm.

## 1 INTRODUCTION

A spatial multiple-choice model uses a scene and instruction to select a candidate. Interpreting accuracy requires three distinct questions: which inputs are needed, whether layout changes alter mean accuracy, and whether they alter answer distributions. Our audits inform spatial-agent assessment without measuring downstream success.

A failed candidate-only prompt does not rule out a shortcut. A frozen question-blind farthest-centroid rule solves 245/326 released STARE 2D instruction questions (75.15%) (Li et al., 2026). On the same 64-question subset, three candidate-only prompts answer 9–11 correctly, versus the rule’s 43.

Crossing scenes and instructions assigns different answers to identical boards, bounding rules that ignore specified inputs. A fixed-rotation obstruction motivates a cyclic construction that balances matched cases while fixing the correct panel. Source-pixel balance needs preprocessing checks: Qwen2.5-VL-32B (Qwen32) inputs admit a perfect query-plus-native-centroid rule. Direct-grid rendering restores seven declared bounds (Figure 1).

Mean accuracy can hide answer changes: Qwen32’s 200 improvements and 197 regressions on 2,048 disk pairs give +0.15 points. Two fresh responses per layout distinguish layout dependence from variation on identical inputs. Our exact categorical permutation algorithm attains the worst tail over missing-response completions.

The balanced construction, preprocessing check and exact test jointly audit specified input use and response stability on the same cases. On three fixed aligned shapes, Qwen32 rejects both omission nulls and response-distribution equality under the response assumptions, without identifying its algorithm.

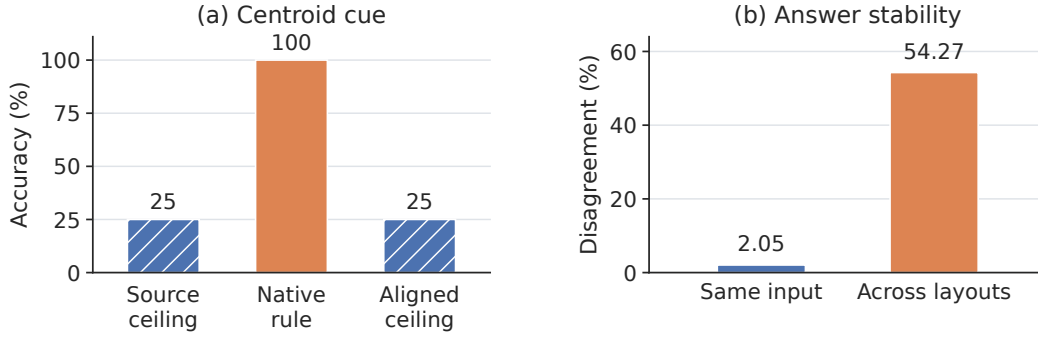


Figure 1: **Centroid cue and response stability.** (a) Hatched blue bars give exact  $1/4$  expected omission ceilings for query plus ordered source or aligned native centroids on 1,024 cases per arm. The solid orange bar shows a perfect post-hoc query-plus-centroid template rule on original native inputs under both grey-mass and threshold descriptors. This audits cue availability, not model use. (b) On 1,024 aligned 78-cell groups, blue shows 42/2,048 (2.05%) same-input and orange 2,223/4,096 (54.27%) cross-layout Qwen32 disagreements.  $\hat{\eta} = 0.5222$  has separate 95% interval  $[0.4586, 0.5859]$ . Inference assumes stable independent kernels; comparisons within groups are correlated. The renderings do not isolate a causal resizing effect.

## 2 RELATED WORK

**Spatial evaluation and visual transformations.** STARE, MindCube and interwoven thinking/drawing motivate spatial assessment (Li et al., 2026; Wang et al., 2026; Wu et al., 2025). Real-3DQA and CVT-Bench examine viewpoint consistency (Ma et al., 2026; Vellamcheti et al., 2026); SpinBench and 3DSRBench include transformation or answer-order controls (Zhang et al., 2026; Ma et al., 2025). Our planar controls test specified input omissions and layout effects; they do not establish 3D generalization. MM-Spatial filters questions using blind-model performance (Daxberger et al., 2025); Section 3 instead gives exact finite-population ceilings for specified omitted-information observables.

**Candidate shortcuts and crossed tests.** Candidate-construction and restricted-input audits expose exploitable cues (Hsieh et al., 2023; Udandarao et al., 2025a; Brown et al., 2025). I-RAVEN identifies context-free answer-set shortcuts and repairs candidate attribute balance (Hu et al., 2021). Winoground uses crossed image-caption pairs and group correctness (Thrush et al., 2022); NoSense/VSC-Repeat studies restricted temporal baselines and label-preserving repetition (Udandarao et al., 2025b). Our audit combines pixel rules, crossed state-instruction questions and fixed-reference, equal-movement comparisons. Balancing the matched reference population in a separate construction puts omission ceilings and paired-mean precision on the same cases. Neither construction identifies an internal strategy.

**Repeated outputs and consistency.** MM-R<sup>3</sup> compares four identical-input repetitions with perturbations (Chou et al., 2025); related work studies repeated language questions (Pinhanez et al., 2026) and adapts LM-head weights for consistency (Chou et al., 2026). We apply classical categorical MMD (Gretton et al., 2012) to the matched inputs. Input use, mean tolerance and response stability retain separate targets and error control. Missing-data MMD inference is established (Zeng et al., 2025); our categorical four-response algorithm computes an attained worst-completion permutation tail.

## 3 METHOD: A JOINT AUDIT ON SHARED CANDIDATES

A rule  $h(C)$  using only the ordered candidate board  $C$  establishes candidate-only shortcut availability. Whether a prompted model finds that rule is a separate question. We audit such rules, then cross inputs and match layouts (Figure 2). O denotes the original layout; S (sham) and I (intervention)

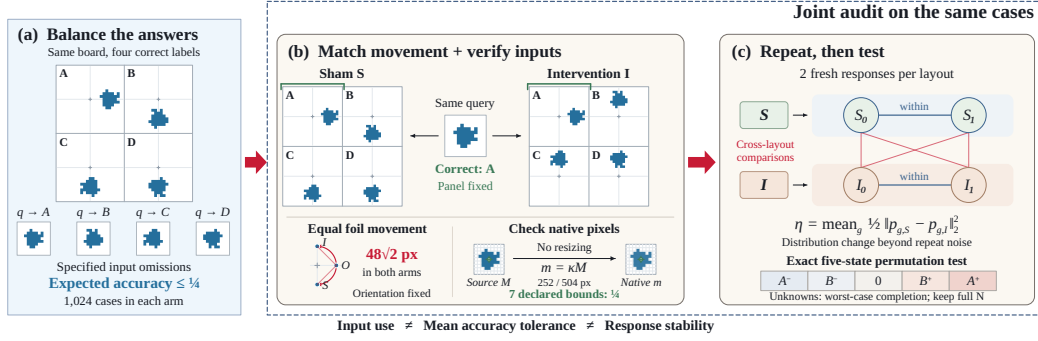


Figure 2: **Joint audit on balanced spatial cases.** (a) Each fixed shape yields 1,024 cases per arm; changing the query assigns all four labels to an identical board. Declared input omissions have  $1/4$  expected ceilings. (b) Foils translate  $48\sqrt{2}$  pixels from O to S or I; the correct panel and orientations stay fixed. Configured Qwen32 processing preserves seven bounds on aligned inputs ( $m = \kappa M$  in the declared foreground region). (c) Two fresh responses per layout contrast within-input and cross-layout agreement under stable independent kernels. The exact five-state permutation test maximizes over unknown completions, retaining all planned groups. Diagrams are schematic; bounds cover declared observables, not arbitrary joint query-and-board rules (Appendix C).

move the incorrect candidates, or foils. Native inputs are the processed images before the visual encoder.

### 3.1 CROSS INPUTS WHILE HOLDING CANDIDATES FIXED

Let  $R$  be a counterclockwise quarter turn and  $U$  an asymmetric centered shape. Cross  $U, RU$  with clockwise/counterclockwise quarter-turn instructions: targets  $R^3U, RU, U, R^2U$  share one randomly permuted board. Omitting either factor leaves two correct labels with identical response marginals, bounding expected quartet accuracy by  $1/2$  ( $1/4$  for candidate-only rules), not realized accuracy. Equal foreground area and centroids at source and explicit 1,344-pixel resolution also bound instruction-aware centroid rules by  $1/2$ ; provider-native learned features are outside this claim. A pixel rotate-and-align solver verifies solvability (Appendix P).

### 3.2 TRANSPORT WRONG CANDIDATES AROUND A FIXED REFERENCE

For a reference designated before acceptance, write  $X_i = U_i + \mathbf{1}b_i^\top$  for the  $n$  indexed points in candidate  $i$ , with  $U_i$  zero-mean and reference centroid  $c \neq 0$ . Transport positions by

$$b_i(\theta) = c + R_\theta(b_i - c), \quad X_i(\theta) = U_i + \mathbf{1}b_i(\theta)^\top. \quad (1)$$

Centered shapes are translated, not rotated. The correct reference panel remains fixed. Pairwise centroid distances and corresponding equal-mass indexed-array distances are invariant:

$$n^{-1} \|X_i(\theta) - X_j(\theta)\|_F^2 = n^{-1} \|U_i - U_j\|_F^2 + \|b_i - b_j\|_2^2. \quad (2)$$

Zero means cancel cross terms; opposite angles move foil  $i$  equally by  $2|\sin(\theta/2)|\|b_i - c\|_2$ . Distance-matrix rules survive without equal perceptual difficulty. Integer translations reverse the reference’s radial rank (Appendix P). Only the reference keeps its correct panel fixed: crossed and matched tests concern distinct question populations.

**Balance the matched population itself.** For ordered centroids  $B = (b_1, \dots, b_K)$  and correct slot  $J$ , fixed  $R \in SO(2) \setminus \{I\}$  sends  $b'_i = Rb_i + (I - R)b_J$ . If  $J$  is uniform conditional separately on  $B$  and  $B'$ , all centroids coincide almost surely (Appendix E). Thus nondegenerate fixed-rotation matching cannot balance both centroid-only arms. This gives no minimum leakage advantage; variable-sign counterexamples prevent applying the theorem to our earlier mixed-sign population.

Our orientation-matching assay instead relaxes common rigid motion. A case is  $(z, r, j)$ :  $z \in \mathbb{Z}_4^4$  has zero sum,  $r \in \mathbb{Z}_4$  is a cyclic orientation offset, and  $j \in \mathbb{Z}_4$  is the correct slot. All phase arithmetic

is modulo four. Using a fixed asymmetric shape, slot  $k$  has orientation  $k + r$  and the centered query has orientation  $q = j + r$ . In arm  $a = +1$  (S) or  $-1$  (I), its position phase is

$$v_k = z_k + a\mathbf{1}\{k \neq j\}, \quad \sum_k v_k = 3a \pmod{4}. \quad (3)$$

Phase  $v_k$  places the shape at one of four radius-48-pixel positions in its panel. The  $64 \cdot 4 \cdot 4 = 1,024$  cases form a complete census. For any fixed visible board  $(r, v)$  and any label  $j$ , the inverse  $z_k = v_k - a\mathbf{1}\{k \neq j\}$  is admissible: that same board occurs with every label once. Fixing query  $q$  and source centroids  $v$  also permits every label, by choosing  $r = q - j$ . Thus candidate-only, query-only and query-plus-source-centroids policies each have expected accuracy at most  $1/4$ . Opposite quarter-turn movements translate the three foils, preserving the correct RGB panel and matching each foil’s movement magnitude. This assay does not preserve inter-candidate distances or vary the instruction. Query plus one candidate orientation already determines the answer; success does not imply comparing every candidate.

**Preserve the bound through preprocessing.** Direct 252/504-pixel grids retain all 1,024 cases for the 78-cell shape and geometry-selected 30/45-cell shapes, with zero cell-coordinate sums, seven-pixel cells and radius 48. Configured preprocessing equals processing without spatial resizing. In the declared region, pre-encoder bfloat16 mass is  $m = \kappa M$  for binary mask  $M$ , with  $\kappa = 134074679/134217728 > 1/2$ : centroids are unchanged and threshold  $m > 1/2$  recovers  $M$ . The cell argument gives  $1/4$  per arm/shape for seven observables: source/native board only, source/native query only, source query plus ordered source centroids, and native query plus ordered grey-mass or threshold centroids (Appendices D.3 and D.5). Functions of each declared observable are covered; arbitrary joint query-and-board rules are not bounded. Aligned cell width and phase radius are  $64/63$  times their nominal prior interpolated values; differing areas also prevent isolating resizing or shape alone.

The earlier one-response-per-layout studies use two-arm Holm omission tests and a separate 95% fixed- $N$  betting interval for mean  $I - S$ , with strict  $\pm 5$ -point containment. Invalid is incorrect; unknowns retain feasible completions. Families are separate by model under stable kernels and independent pairs (Appendix C).

### 3.3 USE SEPARATE TESTS FOR COMPETENCE AND PRACTICAL PRECISION

Crossed competence tests the  $1/2$  group-accuracy ceiling by one-sided Hoeffding under independent groups and specified omitted-input marginals; unknowns minimize accuracy. One 16-test Holm family combines twelve bitmap crossed and four reference McNemar tests. The latter assume fair discordance, not arbitrary heterogeneous mean zero; unknowns maximize  $p$ .

Let  $Y_{g,a} \in \{0, 1\}$  indicate correctness in group  $g$ , arm  $a$ . For expected paired change  $\Delta = N^{-1} \sum_g \mathbb{E}[Y_{g,I} - Y_{g,S}]$ , Bernoulli KL inversion of improvement/regression indicators at  $0.05/16$  gives four simultaneous 95% intervals, allowing heterogeneous means and adverse unknowns. Strict containment in  $(-.05, .05)$  is required; zero discordance on 128 pairs still gives  $\pm 4.41$  points. Precision has a separate family (Appendix O).

**Disk extension and mean precision.** The asymmetric five–seven-disk sampler retains radial-rank reversal, correct-panel and distance invariants; a fixed noncollinearity filter excludes calibration orbits but retains repeated draws. An 80-group core receives quartets; reference pairs cover 2,304 Sol and 2,048 Qwen32 groups (Appendix F). Screening retains positive Sol and negative Qwen32; discovery uses eight Holm tests (six crossed, two reference).

Our fixed-cohort proof adapts bounded-mean betting tools (Waudby-Smith & Ramdas, 2024) to independent heterogeneous means and adverse unknowns (Appendix N). Four tails at 0.0125 give two-model simultaneous 95% coverage. The strict  $\pm 5$ -point margin has no established downstream utility interpretation; the separate-KL comparator cannot replace it. Sol’s post-analysis amendment completes only 855 unattempted tasks, retaining the original stop without retries, new groups or an optional-stopping guarantee (Appendix H).

### 3.4 SEPARATE LAYOUT DEPENDENCE FROM REPEATED-RESPONSE NOISE

Two independent responses per layout keep images, prompts, answer letters and decoding fixed within each repeated input, separating ordinary repeat variation from layout dependence. For group  $g$ , write these as  $S_0, S_1, I_0, I_1$  and define

$$T_g = \frac{1}{2}(\mathbf{1}\{S_0 = S_1\} + \mathbf{1}\{I_0 = I_1\}) - \frac{1}{4} \sum_{a,b=0}^1 \mathbf{1}\{S_a = I_b\}. \quad (4)$$

If repeated responses are independent draws from stable categorical kernels  $p_{g,S}, p_{g,I}$ , then  $\mathbb{E}[T_g] = \eta_g = \frac{1}{2}\|p_{g,S} - p_{g,I}\|_2^2$ . This is one half of the classical unbiased estimator of squared MMD, using the delta kernel (Gretton et al., 2012). The mean  $\eta = N^{-1} \sum_g \eta_g$  is zero exactly when both kernels agree on every planned group. It differs from mean accuracy and from the probability that an individual answer changes causally.

Earlier studies retain invalid as a fifth category and unavailable responses as unknown. Adverse completions of  $T_g \in [-1/2, 1]$  give Hoeffding  $p_H$ ; complete cohorts also allow exact six-assignment  $p_{\text{perm}}$ . They use  $\min(1, 2 \min(p_H, p_{\text{perm}}))$ , setting  $p_{\text{perm}} = 1$  with any unknown, then Holm across two task families. Separate simultaneous  $\eta$  intervals do not reuse the five-point accuracy margin (Appendix K).

**Joint evidence on the balanced population.** Each balanced cohort draws  $S_0, S_1, I_0, I_1$  on all  $N = 1,024$  cases, giving 2,048 responses per arm. For each arm, let  $\bar{Z}_{\text{lower}}$  be the mean of the  $N$  two-replica correctness averages under adverse unknown completions. Each group average lies in  $[0, 1]$ ; each omission  $p$  value is  $\exp\{-2N[\max(0, \bar{Z}_{\text{lower}} - 1/4)]^2\}$ , not a 2,048-trial bound. Under kernel equality, the six assignments within a case are equiprobable. The distribution test maximizes the exact permutation tail over all missing completions, keeping invalid as an observed fifth category. Pooled response multiplicities  $2 + 2$  give  $A_{\pm}$  and  $2 + 1 + 1$  give  $B_{\pm}$ ; all others give zero. Equal pairs separated by arm give  $A_{+}$ ; a within-arm duplicate gives  $B_{+}$ ; otherwise use minus. For  $(A_{-}, B_{-}, 0, B_{+}, A_{+})$ , the centered equiprobable triples of  $4T_{\text{perm}} - 4T_{\text{obs}}$  are  $(0, 0, 6), (0, 0, 3), (0, 0, 0), (-3, -3, 0), (-6, -6, 0)$ . Their pointwise order lets each case choose its greatest attainable state; integer convolution gives the attained worst tail (Appendix L). Completion witnesses are not observations. Each cohort requires its three rejections in its declared Holm family at 0.05: three tests for each 78-cell cohort, six shared tests for the two additional shapes. A separate 95%  $\eta$  interval uses adverse endpoints and radius  $1.5\sqrt{\log(40)/(2N)}$ ; it makes no five-point mean-accuracy claim.

## 4 EXPERIMENTS

### 4.1 SETUP AND COVERAGE

We evaluate ReallLM Gemma3-27B and MistralSmall3.1-24B, local Qwen2.5-VL-3B and Qwen2.5-VL-32B-Instruct (Qwen32), plus Azure GPT-5.6 Sol and Claude Opus 5. Sol and Opus use low and adaptive-high reasoning, respectively, with 8,192-token caps; the other configurations use 2,048. Qwen32 uses greedy bfloat16 generation on one H100 NVL. Unequal settings and native processing preclude equal-compute rankings. The 22,560-request base, 3,072 original repeats, 5,120 initial balanced calls and 16,384 balanced repeats retain 46,978 valid, 118 invalid and 40 unknown. The base has 2,720 groups; original repeats use 384 per model. Balanced studies use 1,024 cases per fixed shape/rendering; Opus uses a prospectively selected 512-case union of four balanced components. Its population is separate. Appendix Table 7 retains study counts, including 10,784 disk requests (5,136 Qwen32; 5,648 Sol).

Initial controls have 96 shape-function and 64 affine groups, split equally between displacement/sign and reflection axis/order. Bitmap symmetry filtering retains 128/138 connected 5–9-tile proposals: 100 rotation orbits and 28 repeats. Four calibration orbits and 240 calls are excluded without accuracy selection. All 7,680 bitmap responses are valid; two pixel oracles solve all 1,536 definitions. Protocols precede formal responses.

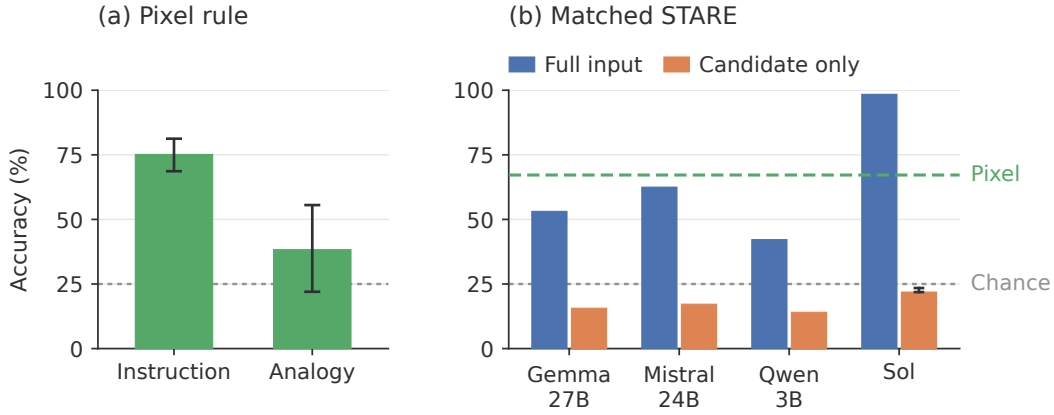


Figure 3: **Public-data controls.** (a) Green bars show the fixed question-blind pixel rule on released instruction tasks (245/326; 196 source groups) and deduplicated analogy transfer (18/47; 32 groups). Whiskers are the existing 95% source-group bootstrap intervals. (b) Blue full-input and orange candidate-only bars use the same 64 STARE questions. The green line is the matched pixel rule (43/64), and grey dashed lines mark 25% chance. Panel b has no sampling intervals; Sol’s candidate-only whisker is the 14–15/64 bound from one unknown response. Denominators remain separate.

#### 4.2 PUBLIC DATA: A SHORTCUT CAN EXIST WHEN A PROMPT FAILS

**A source-grounded observation.** The pinned STARE 2D instruction release (Li et al., 2026) has 333 questions from 200 initial-image hashes. Its inspected generator uses `adjust_to_boundary` to recenter distractors but retains correct-output displacement; the historical generation revision is unverified. Four isolated calibration groups fix detection, ties and failure handling. Removing their seven questions leaves 326 in 196 groups. The rule solves 245/326 (75.15%; group-bootstrap interval 68.66–81.25%), including 9/23 with a failed panel detection under its frozen fallback. The best constant label solves 91/326.

**The same-task prompt comparison.** On the same 64 questions (Figure 3), Gemma, Mistral and Qwen’s full-input/candidate-only gaps survive the six-test Holm family. Sol scores 63/64 versus 14 known correct plus one unknown; this does not identify radial-rule use. Our comparison on photographic MindEdit-Bench (Yu et al., 2026) gives 10/64 versus 7/64, 6/64 versus 15/64, 8/64 versus 10/64, and 32/64 versus 11/64, respectively. The first three gaps do not survive that family; Sol’s later comparison is separate.

**Transfer is limited.** Of 306 visual-analogy rows, 259 duplicate instruction candidate images. Before transfer-label access we freeze the remaining 47 rows from 32 absent case IDs. The unchanged rule solves 18/47 (38.30%; interval 22.00–55.56%), retaining three failures: reliable transfer is unestablished. A descriptive post-result split gives 235/249 instruction translations and 10/77 otherwise. Intervals concern hypothetical group resampling, not census uncertainty.

#### 4.3 RESULT I: CROSSED COMPETENCE EXCLUDES SOME SHORTCUTS, NOT ALL

**A successful crossed test can still admit centroid execution.** Sol solves 123/128 and 121/128 affine A/B questions, with 27/32 and 25/32 fully correct groups; both omission tests reject after 12-test Holm. Other configurations score 23.44–31.25% with no fully correct group. Yet a post-result pixel-centroid executor solves all 352 definitions at both resolutions (704/704 predictions), sufficient without shape-sensitive execution but not identifying Sol’s strategy. Scalar radial and relational-outlier rules each score 32/128 in both crossed families.

**Equal-centroid shapes remove that sufficient explanation.** Table 1 reports every configuration and layout. Sol’s three omission tests reject with Holm-adjusted  $p = 2.90 \times 10^{-20}$ ,  $4.16 \times 10^{-20}$  and  $8.39 \times 10^{-21}$ . This excludes the specified omitted-state, omitted-direction and exact-initial-centroid-only classes under the stated response assumptions; orientation-sensitive features remain

| Model      | Correct / 512 questions |            |              | Fully correct / 128 quartets |
|------------|-------------------------|------------|--------------|------------------------------|
|            | Original                | Sham       | Intervention | O / S / I                    |
| Gemma27B   | 131                     | 124        | 142          | 0 / 0 / 0                    |
| Mistral24B | 143                     | 151        | 147          | 0 / 0 / 0                    |
| Qwen3B     | 125                     | 124        | 122          | 0 / 0 / 0                    |
| Sol        | <b>477</b>              | <b>476</b> | <b>480</b>   | 100 / 97 / 104               |

Table 1: **Bitmap orientation results for four configurations.** O/S/I denote original/sham/intervention;  $K = 4$  counts fully correct quartets among 128 groups per model/layout. Sixteen primary tests share one Holm family. Green/bold marks rejection of the 1/2 omission ceiling, not an equal-compute ranking.

| Model      | S / I correct | W / L   | $\Delta$ (pp) | Simultaneous 95% CI (pp) | Holm $p$ |
|------------|---------------|---------|---------------|--------------------------|----------|
| Gemma27B   | 25 / 38       | 22 / 9  | +10.2         | [-9.3, +28.5]            | 0.383    |
| Mistral24B | 37 / 37       | 13 / 13 | +0.0          | [-18.0, +18.0]           | 1.000    |
| Qwen3B     | 30 / 37       | 8 / 1   | +5.5          | [-5.4, +16.1]            | 0.469    |
| Sol        | 122 / 126     | 5 / 1   | +3.1          | [-6.3, +12.6]            | 1.000    |

Table 2: **Bitmap reference outcomes and precision.** W/L count intervention improvements/regressions on the same 128 questions. Primary Holm  $p$  values and simultaneous KL intervals belong to separate families. Amber intervals extend beyond the frozen  $\pm 5$ -point margin: all four tolerance assessments are inconclusive.

compatible. The other three configurations score 23.8–29.5%, have no fully correct quartet and reject no crossed null; nonrejection does not identify their strategy.

**The disk family reproduces the required-input diagnosis.** In the separate 80-group disk core, Sol scores 300, 300 and 301 of 320 planned O/S/I questions, with three, one and five unknowns. All three conservative omission tests reject after eight-test Holm ( $p = 3.51 \times 10^{-13}$ ,  $3.51 \times 10^{-13}$ ,  $2.58 \times 10^{-13}$ ). Qwen32 scores 56/320, 60/320 and 64/320: no fully correct quartet or rejection.

**A compact orientation feature solves nearly all source tasks.** A post-hoc phase baseline uses pixels and direction; second moments only trigger abstention. For centered complex pixel coordinates  $z_j$ , classical rotation covariance  $m_3(R_\theta X) = e^{3i\theta} m_3(X)$  for  $m_3 = N^{-1} \sum_j z_j^3$  (Flusser, 2000) gives four quarter-turn phases when  $m_3 \neq 0$ . The two-component unit-phase descriptor solves 1,536/1,536 bitmap and 27,612/27,648 disk questions (99.87% for disk), among 29,184 original source tasks: zero wrong predictions and 36 abstentions from three zero-moment disk groups. These are pixel-rule predictions, not extra model calls or evidence of a VLM mechanism (Appendix M). An additional post-hoc source-pixel phase matcher solves all 6,144 S/I questions across the three aligned masks, with zero wrong answers or abstentions (Appendix D.7).

**A perfect reference score still admits a blind alternative.** The relational outlier solves 2,304/2,304 disk S/I references (Qwen32’s 2,048 are nested) and all 80 original-core references, versus 80/320 crossed questions per layout. Bitmap scores are 128/128 references versus 128/512 crossed questions. The radial rule solves all O/S references and none in I. Shared candidates permit strategy switching. The 1,536 valid bitmap candidate-only requests have analytic quartet score 1/4; one response is not four independent trials.

#### 4.4 RESULT II: MEAN TOLERANCE DOES NOT ESTABLISH STABLE ANSWERS

**Earlier comparisons motivate a larger matched cohort.** The 96-group and 32-group affine studies detect no primary matched contrast; the affine interval is  $\pm 48.02$  points. On 128 bitmap pairs, all reference tests and precision assessments are inconclusive (Table 2). Sol’s five improvements and one regression give +3.125 points, with interval  $[-6.26, 12.56]$ .

**A precise average coexists with frequent correctness changes.** Qwen32’s 200 improvements and 197 regressions nearly cancel (Table 3; Figure 4). Its primary interval meets tolerance, although 397/2,048 pairs (19.38%) change correctness and crossed accuracy is 17.5–20.0%. Mean tolerance therefore differs from correctness stability and crossed competence. Sol meets tolerance including

| Model  | Planned /<br>complete pairs | S / I<br>known correct | W / L<br>known | Full-plan $\Delta$<br>(pp) | Primary 95% CI<br>(pp) |
|--------|-----------------------------|------------------------|----------------|----------------------------|------------------------|
| Sol    | 2,304 / 2,278               | 2,168 / 2,155          | 111 / 126      | [-1.22, -0.09]             | [-3.54, 2.23]          |
| Qwen32 | 2,048 / 2,048               | 453 / 456              | 200 / 197      | +0.15                      | [-3.18, 3.48]          |

Table 3: **Disk reference outcomes with full planned denominators.** S/I count known correctness; W/L count improvements/regressions on complete pairs. Sol’s 26 unknown pairs make W/L partial and  $\Delta$  a feasible range; Qwen32 has none. Primary intervals share a separate two-model 95% family. Green means full containment in  $(-5, 5)$  points, not categorical stability.

| Model  | Pairs | S / I<br>correct | W / L     | Invalid /<br>unknown | $\Delta$ (pp) | 95% CI (pp)    |
|--------|-------|------------------|-----------|----------------------|---------------|----------------|
| Sol    | 1,024 | 1023 / 1024      | 1 / 0     | 0 / 0                | +0.10         | [-0.86, +1.08] |
| Qwen32 | 1,024 | 360 / 364        | 125 / 121 | 0 / 0                | +0.39         | [-4.41, +5.20] |
| Opus 5 | 512   | 445 / 444        | 50 / 51   | 101 / 0              | -0.20         | [-6.15, +5.75] |

Table 4: **Balanced omission tests and paired precision share their cases.** Sol/Qwen32 use 1,024 cases; Opus uses a separate balanced 512-case subset. Of 5,120 responses, 101 Opus outputs hit the token cap and count as incorrect; none are unknown. W/L counts improvements/regressions;  $\Delta = I - S$ . Each model has two-arm Holm tests and a separate 95% interval. Green denotes strict unrounded containment in  $(-5, 5)$  points; amber denotes an inconclusive precision result. Neither certifies categorical stability.

all 26 unknown pairs; both reference discovery tests have Holm  $p = 1$ . Only Sol has positive crossed evidence. Separate KL comparators agree; original-stop and post-hoc margin analyses retain the frozen criterion (Appendices I and J).

**A quartet average can hide the matched contrast.** Only the reference keeps its correct panel fixed:  $\Delta_{\text{all}} = (\Delta_{\text{ref}} + 3\Delta_{\text{nonref}})/4$  mixes controls. Qwen32’s reference/nonreference effects  $+5.47/-2.34$  give  $-0.39$  points overall; Sol’s  $+3.125/0$  give  $+0.78125$ . The average cannot replace the matched-reference estimate.

**Competence and mean precision on the same balanced cases.** Sol and Qwen32 reject all three source-omission classes in both arms on the same 1,024 cases (Holm  $p < 10^{-300}$  and  $p \leq 6.70 \times 10^{-10}$ ). Opus also rejects on its separate 512-case population ( $p \leq 3.97 \times 10^{-170}$ ). Only Sol meets mean precision: Qwen32’s interval extends to  $+5.20$  points, and Opus’s spans  $[-6.15, 5.75]$  (Table 4). Qwen32’s 540 letter changes cannot separate layout effects from repeat noise. The cases vary one shape’s orientation and translation; they are not independent shapes.

**Native preprocessing can restore a centroid shortcut.** A post-hoc census of Qwen32’s actual bfloat16 inputs finds a query-plus-native-centroid template rule correct on 1,024/1,024 cases per arm with both grey-mass and threshold-mask descriptors. Source-centroid, native candidate-only and native query-only ceilings remain  $1/4$ . This establishes cue availability, not model use (Appendix D.2).

**Joint evidence on aligned shapes.** All four cohorts reject both omission nulls and kernel equality (Table 5); the two additional masks share a six-test family. Figure 5 contrasts same-input and cross-layout disagreement. Dependence persists on three aligned fixed shapes with native-centroid ceilings of  $1/4$ ; these comparisons identify neither the algorithm nor causal shape, area or resizing effects.

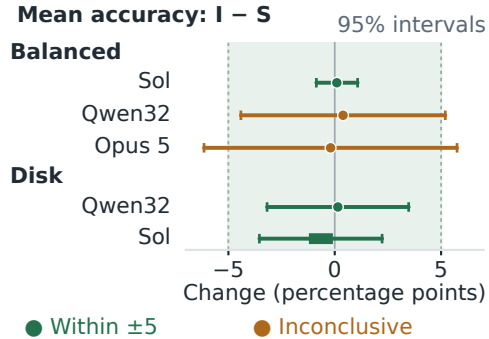


Figure 4: **Mean-change precision.** Whiskers show primary 95% intervals (disk jointly, balanced separately). Shading marks  $\pm 5$  points; green denotes strict interval containment, amber an inconclusive result. Dots show observed  $\Delta$ ; Sol’s thick disk segment spans feasible  $\Delta$  under 26 unknown pairs. Populations and counts: Tables 3 and 4.

| Cohort                         | S / I correct<br>(each / 2,048) | Adjusted $p$<br>S / I; layout                   | $\hat{\eta}$<br>separate 95% CI | Same / cross<br>disagreements | Invalid /<br>unknown |
|--------------------------------|---------------------------------|-------------------------------------------------|---------------------------------|-------------------------------|----------------------|
| Original 78<br>( $N = 1,024$ ) | 722/2,048                       | $4.45 \times 10^{-10} / 2.07 \times 10^{-10}$   | 0.5098<br>[0.4461, 0.5734]      | 41/2,048                      | 0/0                  |
| Aligned 78<br>( $N = 1,024$ )  | 810/2,048                       | $2.21 \times 10^{-19} / 2.21 \times 10^{-19}$   | 0.5222<br>[0.4586, 0.5859]      | 42/2,048                      | 0/0                  |
| Aligned 30<br>( $N = 1,024$ )  | 979/2,048                       | $5.66 \times 10^{-47} / 4.22 \times 10^{-50}$   | 0.4985<br>[0.4349, 0.5622]      | 70/2,048                      | 0/0                  |
| Aligned 45<br>( $N = 1,024$ )  | 1,412/2,048                     | $6.85 \times 10^{-172} / 5.90 \times 10^{-178}$ | 0.2766<br>[0.2130, 0.3403]      | 37/2,048                      | 0/0                  |
|                                | 1,428/2,048                     | $4.55 \times 10^{-136}$                         |                                 | 1,207/4,096                   | 0/0                  |

Table 5: **Joint competence and dependence across fixed shapes and renderings.** Each row: 1,024 groups, 4,096 valid fresh outputs. Holm families are separate for rows 1–2 (three tests each) and shared for rows 3–4 (six);  $\alpha = 0.05$ . Each  $\eta$  interval has separate 95% coverage; no five-point mean claim. Same/cross pools two/four correlated comparisons per group. Original centroid bounds are source-only; aligned bounds cover both native descriptors. Fixed, not randomly sampled, shapes differ in area/rendering.

**Bitmap and disk repeat controls.** Each study uses four fresh calls per group (128 bitmap, 256 disk). Qwen32’s 1,536 answers are valid. Sol has 1,532 valid, one invalid and three unknown (one bitmap, two disk); all calls were attempted.

Qwen32 rejects kernel equality in both families despite unexplained greedy variation. First disk replicas yield 60/256 S and 61/256 I correct (25 improvements, 24 regressions): mean cancellation with dependence.

Sol’s nonrejection leaves invariance inconclusive. Secondary exact worst-completion tests also fail to reject ( $p = 1/3$  and 1; Appendix L), so missing-response fallback alone does not explain this result. Neither study retests the five-point mean-accuracy criterion.

**Qwen32 response disagreement**

$N = 1024$  groups per cohort

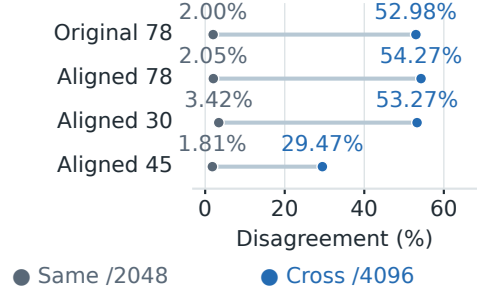


Figure 5: **Repeated-response contrast.** Qwen32: 1,024 groups per row; two same-input and four cross-layout comparisons per group are correlated. Shape, area and rendering effects are not isolated.

| Model / family ( $N$ ) | Same (%)       | Cross (%)      | $\hat{\eta}$       | Holm $p$               | Separate 95% CI  |
|------------------------|----------------|----------------|--------------------|------------------------|------------------|
| Qwen32 / Bitmap (128)  | 4.69           | 44.53          | 0.3984             | $1.03 \times 10^{-25}$ | [0.2022, 0.5947] |
| Qwen32 / Disk (256)    | 4.30           | 41.80          | 0.3750             | $4.91 \times 10^{-45}$ | [0.2362, 0.5138] |
| Sol / Bitmap (128)     | [4.69, 5.08]   | [5.47, 5.86]   | 0.0078             | 1                      | [0.0000, 0.2041] |
| Sol / Disk (256)       | [12.30, 12.70] | [11.33, 11.72] | [-0.0117, -0.0078] | 1                      | [0.0000, 0.1310] |

Table 6: **Same inputs and matched layouts, two responses each.**  $N$  counts groups; same/cross pools two/four correlated comparisons per group. Ranges retain unknown completions; invalid is a fifth category. Holm tests and simultaneous 95%  $\eta$  intervals have separate two-family allocations per model. Blue intensity encodes disagreement, not significance.

## 5 CONCLUSION AND LIMITATIONS

The audits separate input use, mean tolerance and response stability. Transfer, fixed shapes, unequal decoding and perceptual difficulty limit scope. Inference assumes stable independent kernels and groups; internal strategy, agent success and 3D generalization remain unestablished.

## REPRODUCIBILITY STATEMENT

The uploaded supplementary zip file with accompanying code package provides CPU replay entry-points and figure inputs; response-level replay also requires the separately retained frozen scientific bundles. The entrypoints reconstruct stored-output scores and statistics; they do not recreate provider inference.

## AI USE STATEMENT

Generative AI assisted literature search and synthesis, implementation of the stimulus generators and analysis code, and manuscript proofreading.

## REFERENCES

- Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should “train on the test set” to expose exploitable non-visual shortcuts, 2025. arXiv preprint arXiv:2511.04655.
- Shih-Han Chou, Shivam Chandhok, James J. Little, and Leonid Sigal. MM-r<sup>3</sup>: On (in-)consistency of vision-language models (VLMs). In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4762–4788. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.246.
- Shih-Han Chou, Shivam Chandhok, James J. Little, and Leonid Sigal. Test-time consistency in vision language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7789–7798, 2026.
- Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7395–7408, 2025.
- Jan Flusser. On the independence of rotation moment invariants. *Pattern Recognition*, 33(9):1405–1410, 2000. doi: 10.1016/S0031-3203(99)00127-2.
- Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023. Datasets and Benchmarks Track.
- Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1567–1574, 2021. doi: 10.1609/aaai.v35i2.16248.
- Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *International Conference on Learning Representations*, 2026.
- Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, October 2025.
- Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Prisacariu. Do 3d large language models really understand 3d spatial relationships? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 32591–32610, 2026.

- Claudio Pinhanez, Paulo Cavalin, Cassia Sanctos, and Marcelo Carpinette Grave. Small models exhibit limited answer consistency in repetition trials of the multiple-choice MMLU-Redux and MedQA benchmarks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39): 32719–32727, 2026. doi: 10.1609/aaai.v40i39.40550.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.
- Vishaal Udandaraao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good crepe needs more than just sugar: Investigating biases in compositional vision-language benchmarks, 2025a. arXiv preprint arXiv:2506.08227.
- Vishaal Udandaraao, Shyamgopal Karthik, Surabhi S. Nath, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Solving spatial supersensing without spatial supersensing, 2025b. arXiv preprint arXiv:2511.16655.
- Shanmukha Vellamcheti, Uday Kiran Kothapalli, Disharee Bhowmick, and Sathyanarayanan N. Aakur. CVT-Bench: Counterfactual viewpoint transformations reveal unstable spatial representations in multimodal llms, 2026. arXiv preprint arXiv:2603.21114v1.
- Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. In *International Conference on Learning Representations*, 2026.
- Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.
- Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4797.
- Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking object-level counterfactual spatial reasoning in VLMs from in-the-wild photos, 2026. arXiv preprint arXiv:2607.00491.
- Yijin Zeng, Niall Adams, and Dean Bodenham. MMD two-sample testing in the presence of arbitrarily missing data. *Transactions on Machine Learning Research*, 2025.
- Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations*, pp. 70072–70141, 2026.

## Supplementary Material

## What Does Spatial Accuracy Certify? Matched and Crossed Candidate Audits

## FORMAL REQUEST ACCOUNTING

| Study                           | Groups / cases      | Planned slots | Valid  | Invalid | Unknown |
|---------------------------------|---------------------|---------------|--------|---------|---------|
| Public audit                    | 128                 | 1,536         | 1,533  | 2       | 1       |
| Matched layouts                 | 96                  | 1,152         | 1,151  | 0       | 1       |
| Crossed affine                  | 64                  | 1,408         | 1,404  | 4       | 0       |
| Crossed orientation             | 128                 | 7,680         | 7,680  | 0       | 0       |
| Disk (resource amended)         | 2,304               | 10,784        | 10,739 | 10      | 35      |
| Base subtotal                   | 2,720               | 22,560        | 22,507 | 16      | 37      |
| Qwen32 repeats                  | 384 reused          | 1,536         | 1,536  | 0       | 0       |
| Sol repeats                     | 384 reused          | 1,536         | 1,532  | 1       | 3       |
| Balanced Sol                    | 1,024 cases         | 2,048         | 2,048  | 0       | 0       |
| Balanced Qwen32                 | 1,024 cases         | 2,048         | 2,048  | 0       | 0       |
| Balanced Opus5                  | 512 nested          | 1,024         | 923    | 101     | 0       |
| Balanced Qwen32 repeats         | 1,024 reused        | 4,096         | 4,096  | 0       | 0       |
| Native-aligned Qwen32 repeats   | 1,024 new rendering | 4,096         | 4,096  | 0       | 0       |
| Two-shape native Qwen32 repeats | 2,048 cases         | 8,192         | 8,192  | 0       | 0       |
| Combined total                  | —                   | 47,136        | 46,978 | 118     | 40      |

Table 7: **All formal requests with their planned denominators.** The 2,720-group base includes Qwen32’s 2,048 disk groups nested in Sol’s 2,304. The original repeat studies each reuse the same 384 base groups. The original balanced shape supplies 1,024 cases evaluated once per arm in Sol and Qwen32; Opus uses a prospectively selected 512-case subset. Two Qwen32 rendering studies use four fresh responses per case on 1,024 cases each (4,096 calls each). Two additional fixed shapes supply 1,024 cases each, again with four responses per case (8,192 calls total). Overlapping cases and fixed shapes are not pooled as independent groups or datasets. Calibration and CPU pixel-solver predictions are excluded; constructed studies are not independent public benchmark datasets.

## A EXACT RENDERED EXAMPLE OF THE JOINT CONSTRUCTION

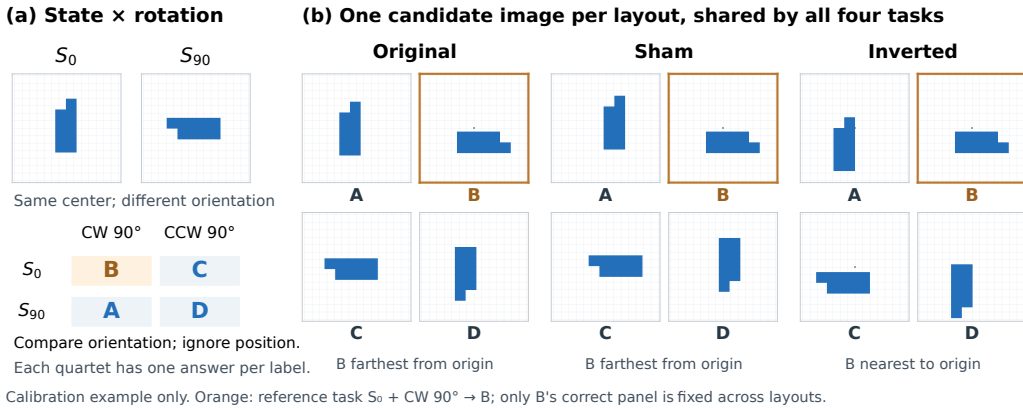


Figure 6: A fixed calibration example from the retained geometry renderer. The initial orientations share area and pixel centroid; the four question combinations use the same candidates. Only the designated reference answer keeps its panel fixed across layouts. Display crops do not replace the original uncropped model inputs. This example was selected independently of formal model outcomes.

## B EXACT DISK-FAMILY INPUTS

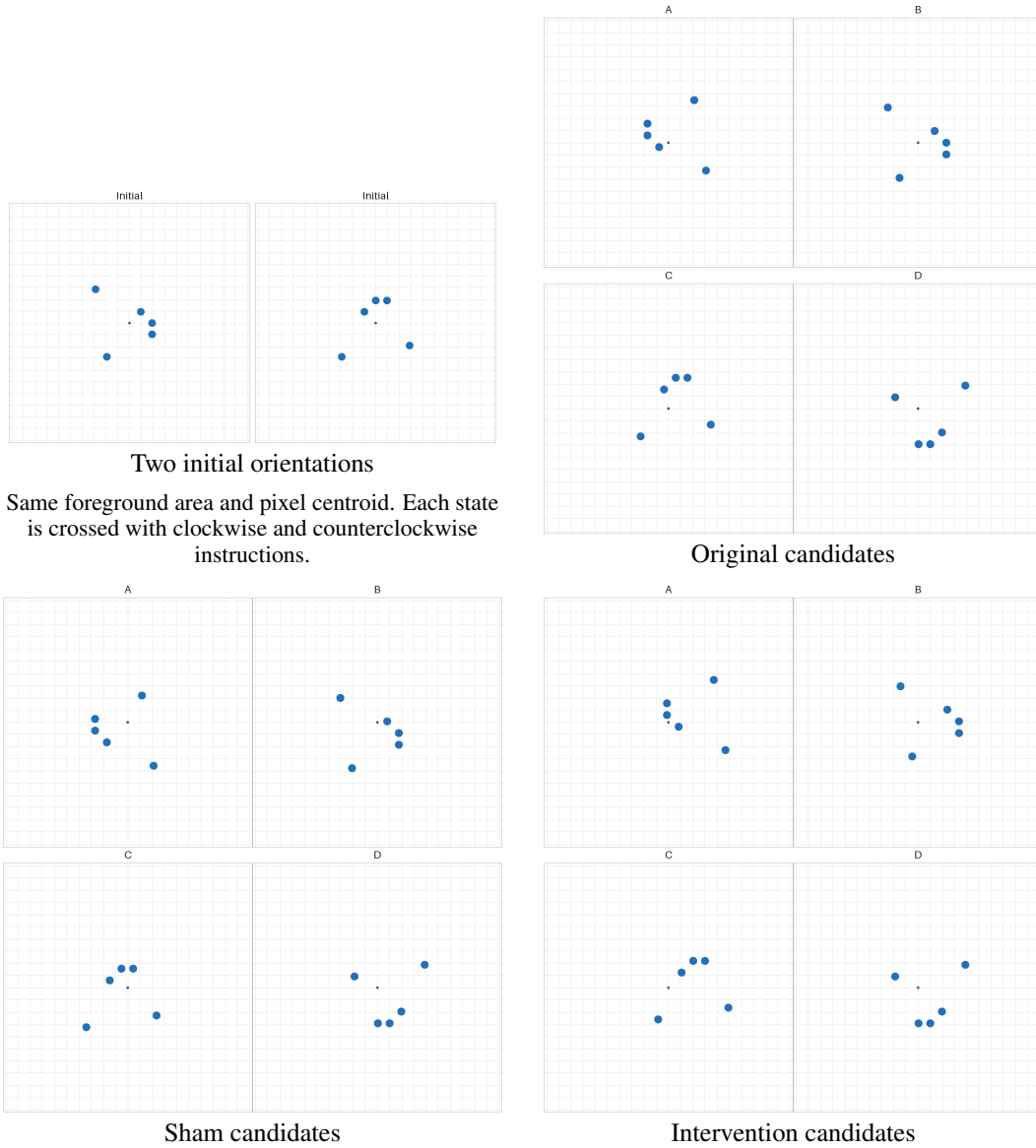
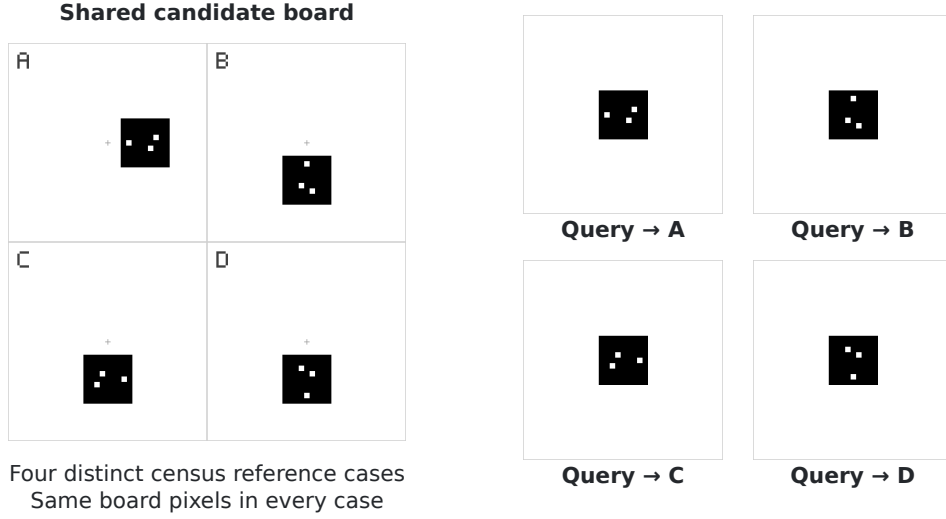
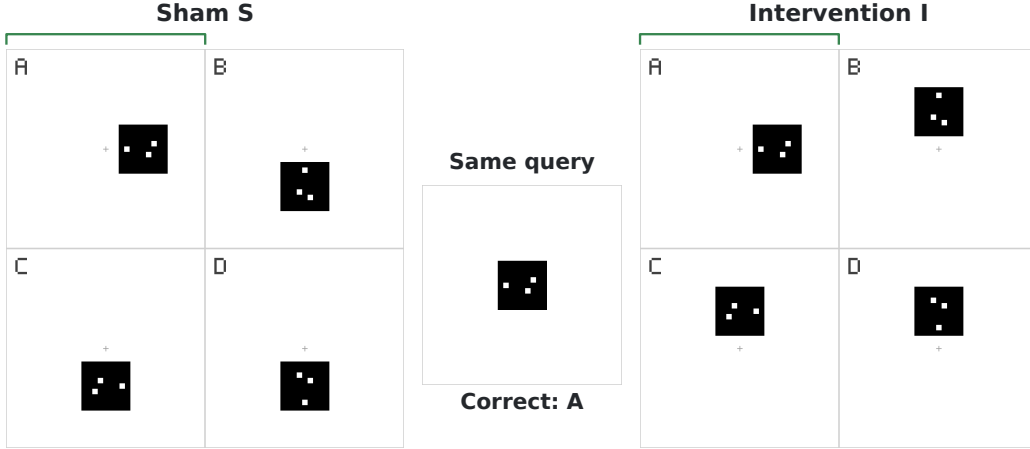


Figure 7: Exact, uncropped model-input PNGs from the first accepted disk-family calibration group, selected before formal model responses. Separated disks jointly form one shape. Candidate D is the designated reference answer and its panel remains unchanged; all four orientation answers remain in the same slots. This is a construction example, not a selected model success or failure. The source group is `constellation_calibration_00000`.

### A One identical S board supports all four answers



### B One matched pair, with a shared query and answer A



Green brackets: the complete correct panel is identical.  
Each foil moves  $48\sqrt{2}$  pixels from O to S and from O to I.

Figure 8: **Exact retained source stimuli from the balanced census.** (A) One identical sham board occurs in four distinct census reference cases, with centered query images whose correct answers are A, B, C and D. The original phase vectors differ; these are not four relabelings of one original case. (B) One of those cases supplies a matched S/I pair with the same query. Green brackets, placed outside the source PNGs, mark the identical complete correct panel. Each foil translates by  $48\sqrt{2}$  pixels from O to S and from O to I; its orientation is unchanged. The source PNGs are embedded without pixel modification. Neither panel shows model outputs.

## C BALANCED CYCLIC POSE CONSTRUCTION AND OMISSION CEILINGS

**Complete finite population.** Use one asymmetric shape with four distinct quarter-turn orientations, indexed by  $\mathbb{Z}_4$ . A case is  $(z, r, j)$ , where  $z \in \mathbb{Z}_4^4$  satisfies  $\sum_k z_k = 0$ ,  $r \in \mathbb{Z}_4$  is a cyclic orientation offset, and  $j \in \mathbb{Z}_4$  is the correct slot (A–D correspond to 0–3). Slot  $k$  has shape orientation  $k + r$ ; the centered query has orientation  $q = j + r$ . There are  $4^3 \cdot 4 \cdot 4 = 1,024$  cases. All arithmetic on these indices is modulo 4. For arm  $a \in \{+1, -1\}$  (S and I respectively), its visible

phases are

$$v_k = z_k + a\mathbf{1}\{k \neq j\}, \quad \sum_k v_k = 3a. \quad (5)$$

Phase  $v$  places the foreground centroid at  $(128, 128) + 48(\cos(\pi v/2), \sin(\pi v/2))$  in its own 256-pixel panel; image coordinates increase downward. The  $512 \times 512$  board uses the ordered top-left panel offsets

$$(t_0, t_1, t_2, t_3) = ((0, 0), (256, 0), (0, 256), (256, 256)) \text{ pixels}, \quad (6)$$

so its ordered panel centers are  $(128, 128), (384, 128), (128, 384), (384, 384)$ . The absolute centroid of slot  $k$  is its panel-local centroid plus  $t_k$ . The renderer uses  $7 \times 7$ -pixel cells with seven-pixel center spacing on the cell grid  $\{-4, \dots, 4\}^2 \setminus \{(-3, 0), (1, 1), (2, -1)\}$ , giving 78 cells and 3,822 black foreground pixels. Their intrinsic centroid is exactly zero in every quarter-turn orientation. These phase changes translate the shape rather than rotating it. The candidate board is denoted  $B(r, v)$ . Distinct phases and orientations remain distinguishable in the retained source pixels, verified by the complete pixel census.

**Why the same board has all four answers.** Fix an arm and any admissible board  $B(r, v)$  with  $\sum_k v_k = 3a$ . For every proposed correct slot  $j$ , define

$$z_k^{(j)} = v_k - a\mathbf{1}\{k \neq j\}. \quad (7)$$

Its sum is zero, so  $(z^{(j)}, r, j)$  is a valid case. Its query orientation is  $q = j + r$ , giving each answer exactly once on this identical board. Conversely, Equation 7 uniquely recovers the original phases from  $(v, j)$ . Thus each of the 256 distinct boards in either arm has four cases, one per correct label. Figure 8A shows actual records sharing one S board.

**Query-only and query-plus-source-centroids cells.** For a fixed query orientation  $q$  and any slot  $j$ , the unique cyclic offset  $r = q - j$ , together with each of the 64 zero-sum original phase vectors, gives 64 cases with answer  $j$ . A query-only cell therefore contains each label 64 times. Next fix both  $q$  and the ordered source-image centroid phases  $v$ . For every  $j$ , choose  $r = q - j$  and recover  $z^{(j)}$  by Equation 7. Each choice is valid and unique. Consequently each query-plus-source-centroids cell also has exactly one case per label, even though only four cyclic orientation permutations are used. There are 256 such cells in each arm.

**Expected-accuracy bound.** Let  $n_{ty}$  count label  $y$  in observable cell  $t$ , and let  $\pi_y(t)$  be a response policy's probability of answering  $y$  when observing  $t$ . A policy depending only on this observable has

$$\frac{1}{N} \sum_t \sum_y n_{ty} \pi_y(t) \leq \frac{1}{N} \sum_t \max_y n_{ty} = \frac{1}{4}, \quad (8)$$

since  $\sum_y \pi_y(t) \leq 1$ ; invalid responses add no correct mass. Always selecting any one label attains  $1/4$ , so the ceiling is exact. The deterministic finite score obeys this bound. For a randomized policy it bounds expected accuracy, not every realized sample. The algebraic ceiling itself does not require independent responses; statistical tests of it additionally assume stable observable-dependent response kernels and independent planned pairs. The centroid class refers to the original source RGB images; interpolation during native preprocessing can introduce other centroid cues and is not covered by that ceiling.

**Controls and limitations.** The 1,024 matched cases have 2,048 distinct full S/I inputs. The correct complete RGB panel is unchanged in every case. Relative to O, all 3,072 foil movements have squared displacement  $48^2 + 48^2 = 4,608$  pixels<sup>2</sup> in each arm, or displacement  $48\sqrt{2}$  pixels. Comparing  $S$  with  $I$ , the six-component vectors of inter-candidate distances change in 768 cases using panel-local centroids, and in all 1,024 cases using absolute board centroids with Equation 6. Neither convention is invariant, and the earlier rigid-transport indexed-array identity does not apply. The instruction is constant, so there is no omitted-instruction theorem. Moreover, the first candidate's orientation reveals  $r$ ; together with the query it gives  $j = q - r$ . Thus query plus one candidate orientation can solve the full population, and the three  $1/4$  ceilings do not establish a need to compare every candidate or identify an internal reasoning algorithm.

**Completed model cohorts and input processing.** Sol (gpt-5.6-sol) and Qwen2.5-VL-32B-Instruct each completed the same 1,024 pairs, with one fresh response per S/I endpoint: 2,048 responses per model. They use the same constant orientation-matching prompt and original  $256 \times 256$  query and  $512 \times 512$  board RGB PNGs. Sol receives the unchanged PNG bytes with high image detail, low reasoning effort, an 8,192-token output cap and two request workers. Qwen32 uses Transformers 4.51.3 on one H100 NVL, bfloat16, SDPA, batch size one, seed zero, greedy decoding with one beam, repetition penalty 1.05, a 2,048-token output cap and a 12,288-token input cap. Its EOS token IDs are 151645 and 151643; the padding ID is 151643. Both collectors forbid retries and impose a four-hour cutoff on new launches; an already active bounded request may finish.

Qwen32 decodes the original RGB images without manual resizing, then applies its frozen native processor with minimum  $56^2$  and maximum  $1344^2$  pixels. Patch size 14 and merge size two produce  $252 \times 252$  and  $504 \times 504$  images, with grids (1, 18, 18) and (1, 36, 36) and 504 total input tokens. A separate CPU census reproduced all 2,048 input tensor fingerprints, verified identical candidate-board tensors across their companion queries, and verified the correct-panel patch subset in all 1,024 S/I pairs. These checks concern the processor inputs, not internal feature or attention invariance. They do not extend the source-centroid theorem to arbitrary centroid cues after native interpolation. Sol’s provider-native processing is not observed. The different processors, output caps and decoding settings make this a comparison of specified model configurations, not an equal-compute or causal model-ranking experiment.

**Parsing and inference scope.** Both parsers accept a decoded JSON object with only the key `answer` and one uppercase A–D value; their retained envelope permits a whole JSON code fence and uses Python’s decoded-object semantics. Sol additionally requires a completed response from the requested model with no refusal; Qwen32 requires a completed generation ending in EOS. An observed incomplete Sol response or returned length-limited/non-EOS Qwen32 generation is invalid even if its text parses. Observed empty, malformed or refusal answers are also invalid and count incorrect. Transport, response-schema or unavailable generation failures remain unknown. Unknowns retain all feasible correctness completions over the planned denominator rather than being dropped. Each model has its own two-arm Holm discovery family and a separate paired-mean precision interval; no joint across-model guarantee is claimed. Greedy decoding and a fixed seed do not prove bitwise determinism, stationary kernels or independence. Appendix D gives execution accounting and the paired contingency counts; Table 4 reports the model results.

## D BALANCED-STUDY EXECUTION AND PAIRED OUTCOMES

**Terminal cohorts.** Sol and Qwen32 each contribute 2,048 responses on the 1,024 matched pairs in Appendix C. Both completed their entire plans with 2,048 valid answers, zero invalid or unknown outcomes, no unattempted endpoints and no retries. Thus the two completed cohorts contribute 4,096 responses. Opus adds 1,024 responses on a prospectively selected union of four balanced components (512 cases): 923 valid and 101 completed invalid, with no unknown, unattempted or retried endpoints. The three cohorts therefore contribute 5,120 responses. Task order, source-image hashes, request/native-input bindings, parsers and analysis rules were frozen before answer collection. Terminal analyses rechecked these bindings and retained all planned pairs. Separate portable CPU replays reproduced all three deterministic scientific output files byte for byte for each model. All 101 Opus invalids (49 S, 52 I) ended with `max_tokens` and empty extracted answer text; the 923 valid outputs ended with `end_turn`. Opus uses adaptive-high reasoning with an 8,192-token output cap. Its subset and configuration differ from the Sol/Qwen32 census; these are separate inferences, not a common-population ranking.

**Zero-intent Qwen32 initialization stop.** The original local launch stopped after model loading but before any request intent or generation. The equality check on the resolved generation configuration failed because Transformers inserted `attn_implementation=sdpa`, while the CPU-prepared configuration omitted that metadata field. SDPA was already the selected attention implementation. The recorded amendment added only this metadata key to the expected resolved configuration. Source PNGs, task order, native tensors, generation parameters, collector, parser and statistical rules were unchanged; no model answer existed on which to condition the amendment. The original closure is retained with zero intents, zero generations and all 2,048 endpoints unat-

| Model  | Both correct | Improvement | Regression | Both wrong | Letter changes |
|--------|--------------|-------------|------------|------------|----------------|
| Sol    | 1,023        | 1           | 0          | 0          | 1              |
| Qwen32 | 239          | 125         | 121        | 539        | 540            |
| Opus 5 | 394          | 50          | 51         | 17         | 27             |

Table 8: **Complete balanced-pair contingencies.** Sol/Qwen32 rows contain 1,024 pairs; Opus contains 512. The four correctness columns partition that row; improvement means S incorrect/I correct and regression means the reverse. Invalid is incorrect. Letter changes use only pairs with two valid A–D answers: 1,024 for each Sol/Qwen32 row and 422 for Opus. They may occur when both answers are wrong. Discovery tests and precision intervals use separate families for each model.

tempted. It supplies no model observations and does not add another 2,048 rows to the completed Qwen32 comparison. The amended cohort alone supplies its 2,048 observed responses and 1,024 paired outcomes.

**Paired outcomes.** Table 8 decomposes each paired mean into its four correctness outcomes. Arm accuracies and precision intervals appear in Table 4. For Qwen32, 125 improvements and 121 regressions leave only four net gains across 1,024 pairs. Its prespecified interval nevertheless extends beyond +5 percentage points, so the frozen strict  $\pm 5$ -point precision criterion is not met; a small observed mean is not evidence of equivalence. Sol meets that criterion. The 540 Qwen32 and one Sol answer-letter disagreements are descriptive: one response per endpoint cannot separate layout dependence from same-input variability. Neither mean criterion certifies response-distribution stability. Opus has 50 improvements and 51 regressions over 512 pairs, giving  $-0.1953125$  points with 95% interval  $[-6.145203921, 5.749464350]$ : its mean-precision assessment is inconclusive. Its 27 letter changes use only the 422 pairs with two valid letters; 90 pairs lack two valid letters. No invalid answer is omitted from accuracy or precision.

#### D.1 FRESH REPEATED RESPONSES ON THE BALANCED POPULATION

The follow-up retains all 1,024 balanced cases and collects four new Qwen32 responses per case, in slots  $S_0, S_1, I_0, I_1$ . No earlier response is reused. The two replicas within each arm have identical prompts, source PNGs, answer letters and native input tensors. The earlier single-response results motivated this follow-up; the new response protocol was fixed before collection, with no selection of cases by their earlier accuracy or disagreement. It uses the same checkpoint, native processor, bfloat16 SDPA and greedy batch-one configuration, with a 2,048-token output cap on one H100 NVL. All 4,096 planned calls ended normally with valid A–D responses; there are no invalid, unknown, unattempted or retried calls. The frozen analysis and its isolated CPU replay agree byte for byte on the three deterministic scientific outputs.

**One joint discovery family.** For each arm, the mean of its two correctness indicators is one bounded observation per case. The omission test therefore uses  $N = 1,024$ , not 2,048 independent trials. Two one-sided Hoeffding tests against the specified  $1/4$  ceiling and the exact categorical test in Appendix L form one three-test Holm family at 0.05. All three must reject for a joint population statement. The adjusted values are  $4.45 \times 10^{-10}$  (S),  $2.07 \times 10^{-10}$  (I), and  $3.25 \times 10^{-251}$  (categorical equality), so this criterion is met. The exact categorical calculation has 518  $A_+$  groups, eight  $B_+$  groups and 498 inactive groups: its unadjusted tail is exactly  $3^{-526}$ . These are population-level exclusions of the specified classes, not evidence that every case uses every candidate or exhibits both effects.

**Separate estimation and response accounting.** The estimated mean half-squared categorical distance is  $\hat{\eta} = 261/512 = 0.509765625$ . A separate two-tail Hoeffding interval, with 0.025 per tail and range  $[-1/2, 1]$ , gives  $[0.4461046164, 0.5734266336]$  after outward numerical rounding. This interval has separate 95% coverage; it is not a joint guarantee with the discovery family and does not retest the five-point mean-accuracy margin. Completed invalid outputs would be one observed fifth category; unknown outputs would retain every compatible completion and the full planned denominator. Stable input-dependent response kernels and independent latent draws within and across

cases are assumptions. Greedy decoding and randomized dispatch do not prove them. The centroid ceiling concerns the exact source images, not arbitrary features after native interpolation.

| Replica | Correct   | Wrong | Invalid | Unknown |
|---------|-----------|-------|---------|---------|
| S0      | 359/1,024 | 665   | 0       | 0       |
| S1      | 363/1,024 | 661   | 0       | 0       |
| I0      | 366/1,024 | 658   | 0       | 0       |
| I1      | 363/1,024 | 661   | 0       | 0       |

Table 9: **Each fresh balanced-repeat slot.** Every slot contains 1,024 planned responses. These descriptive accuracy counts do not define four additional hypothesis tests.

| Pair  | Categorical       | Letter            | Missing |
|-------|-------------------|-------------------|---------|
|       | different / known | different / valid |         |
| S0/S1 | 21/1,024          | 21/1,024          | 0       |
| I0/I1 | 20/1,024          | 20/1,024          | 0       |
| S0/I0 | 548/1,024         | 548/1,024         | 0       |
| S0/I1 | 540/1,024         | 540/1,024         | 0       |
| S1/I0 | 545/1,024         | 545/1,024         | 0       |
| S1/I1 | 537/1,024         | 537/1,024         | 0       |

Table 10: **All six within-case response comparisons.** Same-input disagreement totals 41/2,048 and cross-layout disagreement totals 2,170/4,096. The comparisons share responses and are not independent samples; inference uses the 1,024 cases. All pairs have two valid answers.

## D.2 A SOURCE CENTROID GUARANTEE NEED NOT SURVIVE PREPROCESSING

**Post-hoc diagnostic and observables.** After the original balanced cohort and its fresh-response follow-up, we audited the actual Qwen32 pre-encoder tensors. The processor maps  $256 \times 256$  query images to  $252 \times 252$  and  $512 \times 512$  boards to  $504 \times 504$ . We invert patch packing to reconstruct the normalized RGB pixels. For channel  $c$ , define foreground mass as  $\sigma_c(w_c - x_c)$ , where  $w_c$  is the native bfloat16 white value,  $x_c$  is the native pixel value and  $\sigma_c$  is the retained float32 normalization scale. The first descriptor is the exact rational centroid of the channel-average mass in each shape region. The second thresholds that mass strictly above  $1/2$  and takes the binary-mask centroid. Both exclude letters, borders, the center marker, total mass and area from the supplied centroid features. The query feature is its full native tensor identity. These are specified pre-encoder observables, not centroids of a learned encoder representation.

**Complete census and sufficient solver.** We process all 512 unique boards and four queries. In each arm, both query-plus-centroid observables have 1,024 distinct cells and an exact finite-census accuracy ceiling of one. An explicit rule obtains the orientation templates from the four query silhouettes translated to the known phase positions, then matches the query’s template to the four candidate centroid coordinates. It uses neither the arm identity nor answer labels, has exactly one match for every input, and answers all 1,024 cases correctly in each arm with either descriptor. By contrast, each complete native candidate board still appears once with every label, and each query appears equally often with every label: both omission ceilings remain  $1/4$ .

Across orientations at a fixed phase, the grey-mass centroid separations range from approximately 0.000840 to 0.004396 native pixels, and the threshold-mask separations from 0.008722 to 0.038161 pixels. No tolerance was fitted and no model sensitivity to these differences was measured. We saved the rule and predictions before joining this diagnostic’s answer-key file. However, the authors already knew the construction, label rule and model outcomes. This remains a post-hoc sufficient-solver audit, not a blind or independently held-out experiment.

**Input fidelity and scope.** Foreground-only reprocessing preserves every native shape-region value while removing all non-shape marks. Inverse packing and repacking reproduce all bfloat16 bits for 1,049 tensors, including calibration and translated templates. Reconstructed input segments match all float32 and bfloat16 fingerprints in the 2,048 original and 4,096 fresh-repeat model inputs. This

diagnostic uses CPU preprocessing only and adds no model responses. It shows that the exact source-centroid bound does not transfer to these two native centroid descriptors. It neither refutes source-image balance nor shows that Qwen32 uses this cue. The measured pipeline includes resizing, normalization and bfloat16 conversion; this diagnostic does not isolate which stage causes each difference. The categorical response results remain tests of layout dependence, without identification of an internal mechanism.

### D.3 A SOURCE AND NATIVE GRID ALIGNED RENDERING CONTROL

The observed source-to-native centroid leakage motivated a new rendering control. Its engineering feasibility was assessed after the earlier model results and construction were known, and before preparing a separate response protocol. This chronology does not make the engineering census an independent or untouched confirmatory study. All 1,024 symbolic cases are retained; the construction is not selected using new model responses.

We render the same asymmetric 78-cell shape directly on a  $252 \times 252$  query/panel canvas and a  $504 \times 504$  board, with integer center (126, 126), seven-pixel square cells, and phase radius 48. This changes the earlier 256/512 canvases and (128, 128) center, without resizing their saved images. The cyclic orientations, phase transformations, labels, and reference indices are unchanged. The occupied cell set is  $\{-4, \dots, 4\}^2$  with  $(-3, 0), (1, 1), (2, -1)$  removed. Each shape contains  $78 \cdot 49 = 3,822$  black pixels. Its four quarter-turn orientations have the same centroid because the occupied cell coordinates sum to zero. In pixel-center coordinates, the query centroid is  $(253/2, 253/2)$ ; candidate centroids add the four radius-48 axis displacements.

The new dimensions are multiples of the processor’s 28-pixel patch/merge factor. On all 772 distinct images (768 original/sham/intervention boards and four queries), configured preprocessing produces the same float32 patch bytes as preprocessing with spatial resizing disabled. We reconstruct the two identical static frames from the actual bfloat16 patch representation and verify exact unpack–repack bit equality. The declared shape region uses pixel indices  $[40, 212]^2$ , with the fixed center-marker region  $[116, 136]^2$  excluded on boards. Isolated-foreground copies verify that this region retains every shape pixel and contains no letters, border, or other non-shape mass.

Let  $B(x, y)$  be the binary source shape mask in that region, and let  $b_c, w_c$  be the bfloat16 normalized black and white values in channel  $c$ . Absence of spatial resampling gives  $z_c = b_c B + w_c(1 - B)$ . With the normalization scales  $\sigma_c$  interpreted as their exact float32 values, the declared grey mass is

$$m(x, y) = \frac{1}{3} \sum_c \sigma_c (w_c - z_c(x, y)) = \frac{134074679}{134217728} B(x, y) = \kappa B(x, y). \quad (9)$$

Indeed,  $2^{24}(w_c - b_c) = (62390272, 64225280, 60686336)$  and  $2^{23}\sigma_c = (2253428, 2191965, 2313386)$ ; their dot product divided by  $3 \cdot 2^{47}$  gives the displayed fraction. Thus  $\kappa > 1/2$  cancels from centroid ratios, and the strict threshold  $m > 1/2$  recovers  $B$  exactly. All 3,076 panels satisfy this identity and exact equality of source, native grey-mass, and native threshold centroids. The full census also checks 1,024 unchanged correct panels in source and native bytes, 1,024 zero correct movements, and 3,072 foils with equal squared displacement 4,608 under the two transformations. Native preprocessing is additionally regenerated for all 2,048 distinct case–arm inputs; the two planned replicas per arm share their entire input fingerprints.

For each arm separately, each of seven observables has maximum finite-census accuracy  $256/1024 = 1/4$ : source board only, native board only, source query only, native query only, query plus ordered source centroids, query plus ordered native grey-mass centroids, and query plus ordered native threshold centroids. The source-centroid observable uses the source query, and the two native centroid observables use the native query. Each query-only observable has four cells of 256 cases; each other observable has 256 cells of four cases. Every cell contains equally many A–D answers. Here a native observable refers to the declared pre-encoder bfloat16 input or descriptor. Functions of any declared observable remain covered; arbitrary functions of the joint query and board are not bounded. The new native cell size and phase radius are 64/63 times their nominal earlier interpolated values, so this control does not isolate the causal effect of resizing alone.

#### D.4 FRESH RESPONSES AFTER NATIVE ALIGNMENT

After the complete input census, a separate protocol fixed four fresh Qwen32 responses for every one of the 1,024 aligned 78-cell cases. The checkpoint, bfloat16 SDPA, greedy batch-one decoding and 2,048-token cap match the original repeat study. The 4,096 requests are shuffled globally with fixed seed 20260921252504; each same-arm pair retains identical prompts, source images and native tensors. All 4,096 responses are valid A-D answers, with zero invalid, unknown, unattempted or retried slots. Collection took 2,955.41 seconds on one H100 NVL. The three scientific outputs of the closed analyzer and isolated CPU replay agree byte for byte.

Correct counts in slots  $(S_0, S_1, I_0, I_1)$  are (404, 406, 407, 404) out of 1,024 each. The two omission tests use the case-level replica means and the union of the seven declared observable classes, each with mean ceiling  $1/4$ . They share one three-test Holm family with categorical equality, independently of the earlier study’s error family. All three reject: the two omission adjusted values are both  $2.21 \times 10^{-19}$  (rounded upward), and the categorical value is  $2.04 \times 10^{-257}$  (rounded upward). The exact calculation contains 529  $A_+$ , twelve  $B_+$ , one  $B_-$  and 482 inactive cases. Its unadjusted tail is  $3^{-539}$ : every  $A$  must take its positive assignment and at least twelve of thirteen  $B$  cases must do so.

The observed mean is  $\hat{\eta} = 2139/4096 = 0.522216796875$ , with separate 95% interval  $[0.4585557883, 0.5858778054]$  under the same bounded-score assumptions. Same-input disagreement counts are 18 for  $S_0/S_1$  and 24 for  $I_0/I_1$ . The four cross-layout counts for  $S_0/I_0$ ,  $S_0/I_1$ ,  $S_1/I_0$  and  $S_1/I_1$  are 557, 553, 559 and 554, each out of 1,024. These six correlated comparisons supply descriptive counts, not six additional tests. This study does not evaluate the five-point mean-accuracy criterion. The native-footprint difference prevents an isolated causal attribution to resizing; joint rejection holds on a rendering whose two declared query-plus-native-centroid classes have a  $1/4$  ceiling.

#### D.5 TWO ADDITIONAL FIXED SILHOUETTES ON THE NATIVE-ALIGNED GRID

The two additional silhouettes extend the rendering control of Appendix D.3 while retaining its cyclic case family, phase transformations, canvas dimensions and processor. They were selected using integer geometry before responses from their own experiment were collected. The earlier construction, model studies and preprocessing limitation were already known; this is an author-designed extension with a recorded selection rule. The occupied areas differ, so the extension is not an isolated causal contrast of shape morphology.

**Deterministic mask selection.** Start from the centered  $3 \times 3$  core  $\{-1, 0, 1\}^2$  inside  $\{-4, \dots, 4\}^2$ . At each growth step, enumerate unordered triples of unoccupied cells whose coordinate vectors sum to zero. Sort these triples lexicographically, shuffle them using the shape’s seeded Python Random instance, and retain the first addition that keeps the foreground four-connected and its empty complement four-connected in  $\{-5, \dots, 5\}^2$ . Stop at 30 or 45 occupied cells. Retain the first complete candidate with zero first moments, no holes, eight distinct dihedral images and a nonzero third central-moment tensor. Seeds 202609212730 and 202609212745 produce the two masks in Figure 9; both first candidates pass, with no rejected complete masks. Every occupied cell and growth step is recorded in `selected_masks.json` in the portable census material. The differing cell counts exclude translation, rotation or reflection equivalence to each other or to the earlier 78-cell mask.

| Mask      | Cells | Black pixels | $\sum(x^3, x^2y, xy^2, y^3)$ |
|-----------|-------|--------------|------------------------------|
| triplet30 | 30    | 1,470        | (36, 27, -19, -18)           |
| triplet45 | 45    | 2,205        | (6, -47, 5, 66)              |

Here  $(x, y)$  are integer cell-center coordinates relative to the centroid. The sums of  $x$  and  $y$  are exactly zero. Rasterizing equal seven-pixel squares preserves the centroid; normalized third moments are multiplied by  $7^3$ , because the symmetric within-cell terms multiply vanishing first moments.

**Exact source and native checks.** Each query/panel is  $252 \times 252$  and each board is  $504 \times 504$ , with integer center (126, 126) and radius-48 phase displacements. For each shape, all  $64 \cdot 4 \cdot$

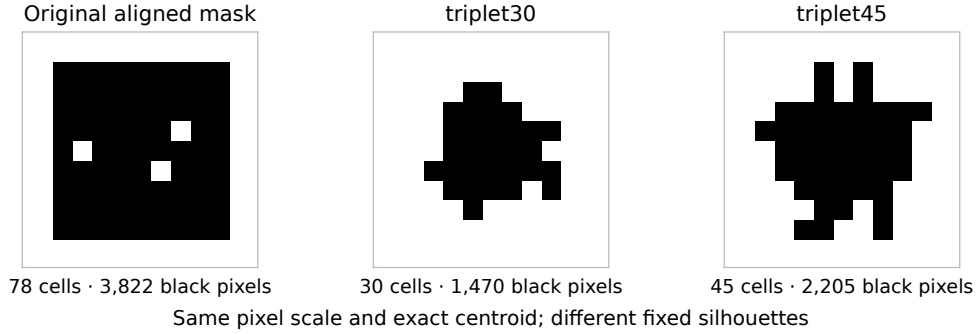


Figure 9: **Three fixed silhouettes with the same exact centroid.** Unresampled central  $84 \times 84$  crops from the saved 252-pixel source queries, shown at a common scale and orientation. Every occupied cell is  $7 \times 7$  pixels; each full query has pixel-center centroid  $(253/2, 253/2)$ . The 30- and 45-cell masks are connected and hole-free. The native census separately verifies preservation of these masks in the declared bfloat16 descriptors; the displayed panels are source RGB crops.

4 = 1,024 symbolic cases are retained. Their 772 distinct source images comprise 768 original/sham/intervention boards and four queries. The common region  $[40, 212]^2$ , with  $[116, 136]^2$  excluded on boards, retains every shape pixel and excludes letters, border and marker, as checked against isolated-foreground images. Configured preprocessing equals preprocessing without spatial resizing in exact float32 patch bytes. Unpacking and repacking the bfloat16 representation preserves every bit and both static frames. Equation 9 applies unchanged:  $m = (134074679/134217728)B$  in the declared region. Consequently, source, native grey-mass and native threshold centroids coincide exactly for all 3,076 panels per shape, with no tolerance-based grouping.

For each shape and arm, the seven observables listed in the main text have maximum finite-census accuracy  $256/1024 = 1/4$ . Each query-only partition has four cells of 256 cases, containing 64 occurrences of each A–D answer. Each of the other five partitions has 256 cells of four cases, containing each answer once. A stable randomized policy measurable in any one of these observables therefore has mean accuracy at most  $1/4$  over the uniform census; the union of the seven whole policy classes shares that bound. It does not allow switching between observables using extra information or jointly observing the full board and query.

The two censuses together verify 1,544 images and 6,152 panels, 2,048 fixed correct panels in both source and native bytes, 2,048 zero correct movements, and 6,144 foil movements with squared displacement 4,608 under each transformation. They also regenerate all 4,096 distinct case–arm processor inputs and compare 8,192 prepared replica records, including rendered prompts, token IDs, attention masks, grids and float32/bfloat16 pixel fingerprints. These are input checks, not model response counts. All statements concern these fixed masks and specified pre-encoder observables. The 64/63 footprint qualification in Appendix D.3 also applies here.

**Portable CPU reproduction.** Preserve the project-relative paths in `portable_sources.json` and use the dependencies and preprocessing-source hashes recorded in `environment.json`. From the evidence root, with an output path that does not already exist, run

```
python3 -B \
  v9/native_aligned_two_shapes-20260921/paper_support/recount.py \
  --root . --output new-recount.json
```

The command verifies the selected masks, every source panel, all seven label partitions and the complete native input fingerprints against `recount-result.json`. It uses two CPU threads and shared, hash-bound processor metadata with local-only loading and remote code disabled. It loads no model weights and performs no inference.

## D.6 FRESH RESPONSES ON TWO ADDITIONAL FIXED SHAPES

The two masks in Appendix D.5 each retain all 1,024 cyclic cases. A separate protocol specifies four fresh Qwen32 responses per case,  $S_0, S_1, I_0, I_1$ : 4,096 requests per shape and 8,192 in one global shuffle (seed 20260921252819). No earlier response or outcome-based case selection enters this collection. Same-arm replicas have identical prompts, source images and native input tensors. Decoding retains bfloat16 SDPA, greedy batch-one generation, a 2,048-token cap and repetition penalty 1.05; the aligned 252/504-pixel inputs require no manual or native spatial resizing. The collection limit is four hours on one GPU. The first infrastructure unknown stops the shuffled stream without retries; every unattempted slot remains unknown in its original shape population.

The two omission tests and categorical test for each shape form one *six-test* Holm family at 0.05, spanning both shapes. Each omission bound uses 1,024 two-replica group means, not 2,048 independent trials. A shape-level conjunction requires its three rejections within that shared family; a two-shape conjunction requires all six. Invalid is the fifth observed category, while the exact permutation test maximizes its upper tail over all missing completions (Appendix L). Each shape has a separate 95% interval for  $\eta$ ; neither simultaneous coverage across shapes nor a five-point mean-accuracy claim is made. Stable independent kernels and groups remain assumptions, not consequences of greedy decoding. The seven declared omission bounds cover functions of each observable; arbitrary functions of the joint query and board are not bounded. The masks are fixed, differ in occupied area, and do not constitute random-shape sampling or an isolated shape/area or resizing contrast.

**Closed response accounting and decisions.** The 8,192 planned slots contain 8,192 valid, 0 completed invalid and 0 unknown responses, including 0 unattempted and 0 unresolved-intent slots. Collection stopped after all planned requests had been attempted, taking 5,757.86 seconds. No calls were retried. The frozen analyzer and isolated CPU replay agree byte for byte on all three scientific outputs. A separate implementation verifies completion supports, the exact tail and global Holm correction.

All six tests reject; both fixed shape populations meet their three-test criteria within the shared family.

| Shape    | Holm S                  | Holm I                  | Holm categorical        | Statistic | Separate 95% $\eta$ CI |
|----------|-------------------------|-------------------------|-------------------------|-----------|------------------------|
| 30 cells | $5.66 \times 10^{-47}$  | $4.22 \times 10^{-50}$  | $8.77 \times 10^{-244}$ | 1021/2048 | [0.434874, 0.562197]   |
| 45 cells | $6.85 \times 10^{-172}$ | $5.90 \times 10^{-178}$ | $4.55 \times 10^{-136}$ | 1133/4096 | [0.212950, 0.340273]   |

Table 11: Global discovery and separate estimation on the two fixed shapes. Both rows belong to one six-test Holm family. Adjusted  $p$  bounds are rounded upward; each  $\eta$  interval is separately 95%, not simultaneous across shapes or with discovery. The statistic is an exact complete-sample mean or a completion range; interval endpoints round outward. No five-point mean-accuracy test is made.

| Shape    | Slot  | Correct | Valid wrong | Invalid | Unknown |
|----------|-------|---------|-------------|---------|---------|
| 30 cells | $S_0$ | 490     | 534         | 0       | 0       |
| 30 cells | $S_1$ | 489     | 535         | 0       | 0       |
| 30 cells | $I_0$ | 500     | 524         | 0       | 0       |
| 30 cells | $I_1$ | 496     | 528         | 0       | 0       |
| 45 cells | $S_0$ | 708     | 316         | 0       | 0       |
| 45 cells | $S_1$ | 704     | 320         | 0       | 0       |
| 45 cells | $I_0$ | 716     | 308         | 0       | 0       |
| 45 cells | $I_1$ | 712     | 312         | 0       | 0       |

Table 12: Every replica on each fixed shape. Each row retains 1,024 planned outputs. Correctness is descriptive and these eight rows are not eight additional tests.

The exact raw tail is  $\Pr(2U_A + U_B \geq 2n_{A_+} + n_{B_+})$ , where  $U_A \sim \text{Bin}(n_{A_-} + n_{A_+}, 1/3)$  and  $U_B \sim \text{Bin}(n_{B_-} + n_{B_+}, 1/3)$  are independent. The inputs and source/native omission census are specified in Appendix D.5; the partial-completion method is in Appendix L. These results do not identify per-case conjunction, all-candidate reasoning, model use of the available phase rule, or broad shape and agent generalization.

| Shape    | Pair      | Known cat. | Cat. disagree | Both letters | Letter disagree | Missing |
|----------|-----------|------------|---------------|--------------|-----------------|---------|
| 30 cells | $S_0/S_1$ | 1,024      | 33            | 1,024        | 33              | 0       |
| 30 cells | $I_0/I_1$ | 1,024      | 37            | 1,024        | 37              | 0       |
| 30 cells | $S_0/I_0$ | 1,024      | 549           | 1,024        | 549             | 0       |
| 30 cells | $S_0/I_1$ | 1,024      | 546           | 1,024        | 546             | 0       |
| 30 cells | $S_1/I_0$ | 1,024      | 545           | 1,024        | 545             | 0       |
| 30 cells | $S_1/I_1$ | 1,024      | 542           | 1,024        | 542             | 0       |
| 45 cells | $S_0/S_1$ | 1,024      | 15            | 1,024        | 15              | 0       |
| 45 cells | $I_0/I_1$ | 1,024      | 22            | 1,024        | 22              | 0       |
| 45 cells | $S_0/I_0$ | 1,024      | 300           | 1,024        | 300             | 0       |
| 45 cells | $S_0/I_1$ | 1,024      | 299           | 1,024        | 299             | 0       |
| 45 cells | $S_1/I_0$ | 1,024      | 304           | 1,024        | 304             | 0       |
| 45 cells | $S_1/I_1$ | 1,024      | 304           | 1,024        | 304             | 0       |

Table 13: All six within-case comparisons per shape. Each row has 1,024 planned pairs. Known categorical pairs include INVALID; letter pairs require two valid A–D outputs. Missing pairs remain separate. The comparisons overlap and are correlated, not additional independent samples.

| Shape    | $A_-$ | $B_-$ | 0   | $B_+$ | $A_+$ |
|----------|-------|-------|-----|-------|-------|
| 30 cells | 2     | 0     | 500 | 21    | 501   |
| 45 cells | 0     | 1     | 737 | 5     | 281   |

Table 14: Five-state exact categorical accounting, with 1,024 cases per shape. If any output is unknown, these are attained worst-completion states, not observed imputed answers.

#### D.7 A SOURCE-PIXEL PHASE SOLVER FOR THE THREE ALIGNED SILHOUETTES

An additional post-hoc baseline applies classical third-moment phase matching to the source RGB inputs of all three fixed aligned masks. The authors already knew the construction, answer rule and earlier results. Saving code and predictions before a separate key join records provenance; it does not make the study blind. The predictor receives only the query and candidate RGB arrays. Fixed framing and foreground extraction use the declared shape region; image paths, case identifiers and answer keys never enter prediction.

For mathematical complex pixel coordinates  $p_j = x_j - iy_j$ , let  $S_k = \sum_j p_j^k$ . The exact integer expansion

$$C_3 = N^2 S_3 - 3N S_1 S_2 + 2S_1^3 = N^3 m_3$$

removes translation without floating cancellation. Dividing the real/imaginary integer pair by its greatest common divisor represents its signed phase. The query’s phase is matched to the four candidates. Invalid or empty foreground, zero moments, ties and absent matches cause retained abstentions; there is no fallback solver or fitted parameter. This adapts the classical identity from Appendix M to black foreground and direct query matching; the earlier executable is not run unchanged.

Every shape has 1,024/1,024 correct in each S and I arm: 6,144/6,144 in total, with zero wrong answers or abstentions. The auxiliary O census is separately 3,072/3,072. These are overlapping constructed input questions, not additional VLM calls or independent shape draws. All 2,316 source PNGs are checked; 9,216 pre-join predictions and joined outcomes are retained. Replaying the frozen scripts from a fresh copy reproduces both prediction and joined-output files byte for byte. The scripts are `predict.py` and `score.py` under `v9/aligned_phase_baseline-20260921/`; their protocol, source-image inventory and accompanying proof give the exact dependency paths.

Asymmetric shapes can have zero third moment; unrelated shapes can share phases, and arbitrary-angle rotation has a 120-degree phase ambiguity. The result is therefore restricted to these nonzero-moment quarter-turn inputs. No native baseline or encoder is executed here. Conditional on the separately verified identity  $m = \kappa M$ , normalized weighted third moments also coincide because  $\kappa$  cancels, and the half-threshold recovers the source mask. This is an analytic implication for the declared pre-encoder descriptors, not a measured native accuracy or a model-mechanism claim.

## E A FIXED-ROTATION OBSTRUCTION AND ITS BOUNDARIES

**Statement.** Let  $K \geq 2$ , let  $B = (b_1, \dots, b_K)$  be a random ordered board in  $(\mathbb{R}^2)^K$ , and let  $J \in \{1, \dots, K\}$ . Fix one deterministic  $R \in SO(2) \setminus \{I\}$ , shared by all samples, and define  $b'_i = Rb_i + (I - R)b_J$ . If

$$\Pr(J = j \mid B) = \Pr(J = j \mid B') = 1/K \quad \text{a.s. for every } j, \quad (10)$$

then  $b_1 = \dots = b_K$  almost surely. No moment assumption is needed. Conversely, coincident centroids give  $B' = B$ , so conditional uniformity on  $B$  also holds on  $B'$ . For a finite label set, optimal board-only accuracy is  $\mathbb{E}[\max_j \Pr(J = j \mid B)]$ , which equals  $1/K$  exactly when the posterior is uniform. This explains the accuracy interpretation of the statement.

**Proof.** Let  $T_j$  be the invertible board map that rotates about candidate  $j$ , and let  $\mu = \text{Law}(B)$ ,  $\nu = \text{Law}(B')$ . Equation 10 implies  $(T_j)_\# \mu = \nu$  for every  $j$ , where  $\#$  denotes the distribution after applying a map. Thus  $\mu$  is invariant under  $T_k^{-1}T_j$ . Direct composition gives the common translation

$$(T_k^{-1}T_j(B))_i = b_i + v(B), \quad v(B) = (I - R)(b_j - b_k). \quad (11)$$

This translation preserves its own offset  $v(B)$ . On the invariant set  $H = \{B : v(B) \neq 0\}$ , put  $z(B) = \langle m(B), v(B) \rangle / \|v(B)\|^2$ , where  $m(B) = K^{-1} \sum_i b_i$ . The map increases  $z$  by exactly one. The disjoint measurable sets  $E_n = \{B \in H : n \leq z(B) < n + 1\}$ ,  $n \in \mathbb{Z}$ , partition  $H$  and are mapped successively onto one another. Invariance forces their probabilities to be equal. Finite total probability and countable additivity imply that each has probability zero, and hence  $\mu(H) = 0$ . Since  $I - R$  is invertible in two dimensions,  $b_j = b_k$  almost surely. There are finitely many pairs, proving the result.

**A short second-moment interpretation.** For finite second moments, write  $s(B) = K^{-1} \sum_i \|b_i - m(B)\|^2$  and  $c = 2 - 2 \cos \theta > 0$ . The transform preserves  $s$  and has  $m(B') = m(B) + (I - R)(b_J - m(B))$ . Uniformity conditional on  $B$  therefore gives

$$\mathbb{E}\|m(B')\|^2 - \mathbb{E}\|m(B)\|^2 = c \mathbb{E}s(B). \quad (12)$$

Applying the same calculation to the inverse transform under uniformity conditional on  $B'$  reverses the left side and preserves the right side. Positive expected scatter would require both differences to be positive.

**Why a varying rotation invalidates the conclusion.** For  $K = 2$ , take  $a = (0, 0)$ ,  $b = (1, 0)$ ,  $c = (1, 1)$  and  $d = (0, 1)$ . Draw  $B$  uniformly from  $\{(a, b), (c, d)\}$  and draw  $J$  independently uniformly. Rotate about candidate  $J$  by  $+90^\circ$  for  $J = 1$  and  $-90^\circ$  for  $J = 2$ . For  $B = (a, b)$  the two outputs are  $(a, d)$  and  $(c, b)$ ; for  $B = (c, d)$  they are  $(c, b)$  and  $(a, d)$ , respectively. Each input and each output board has both labels equally often, despite positive scatter. Adding slot-swapped copies even makes rotation sign marginally independent of both the input board and the label. It still does not give conditional label uniformity given board *and* sign.

The same issue exists for arbitrary angle magnitude  $\theta \in (0, \pi)$ . In complex coordinates let  $\alpha = \pi - \theta$ , draw  $t$  uniformly on the circle, use  $B = (e^{it}, e^{i(t+\alpha)})$ , and draw  $J$  independently uniformly. Rotations  $+\theta$  for  $J = 1$  and  $-\theta$  for  $J = 2$  give  $(e^{it}, e^{i(t-\alpha)})$  and  $(e^{i(t+2\alpha)}, e^{i(t+\alpha)})$ . The ordered output laws agree by a shift of uniform  $t$ . This is a continuous population construction, not a finite rendered census.

**No uniform effect-size guarantee.** Take  $K = 2$ ,  $R = -I$ , and  $B_m = ((m, 0), (m + 1, 0))$  for uniform  $m \in \{0, \dots, N - 1\}$ ,  $N \geq 2$ , with independent uniform  $J$ . The input accuracy ceiling is  $1/2$  and scatter is always  $1/4$ . The two output boards are  $((m, 0), (m - 1, 0))$  and  $((m + 2, 0), (m + 1, 0))$ . Their index ranges overlap at  $N - 2$  positions and have four single-label boundary positions. Their exact Bayes accuracy is  $1/2 + 1/N$ , approaching chance. Thus the obstruction supplies neither a distribution-free positive margin nor a claim that a tested model uses the available cue.

**Scope for the actual assays.** The theorem observes ordered absolute centroid coordinates. Centering away the board mean, clipping, nonrigid moves, changing slots, or varying the rotation

changes its hypotheses. Our earlier transports use opposite signed angles; their relative rotation is  $R_{\text{rel}} = -(7/25)I - (24s/25)J_2$ , where  $s \in \{-1, 1\}$  varies by case and  $J_2$  is the  $90^\circ$  rotation matrix. The theorem can be applied within a fixed-sign stratum only if both conditional balance requirements hold in that stratum. It does not imply leakage in their pooled population; the separately executed outlier rule establishes that empirical result. The cyclic construction instead relaxes common rigid motion and verifies its exact ceilings directly. Finally, empirical Bayes accuracy on a finite census is a statement about that census: memorizing unique boards need not generalize to unseen boards.

## F DISK-FAMILY SAMPLING, TRANSPORT AND COHORTS

The disk family uses a separate analytic shape sampler rather than a connected tile growth process. It tests the same two inferential questions with different shape support and a noncollinear candidate layout. This is a controlled planar rendering family, not an additional public natural-image benchmark.

**Shapes and accepted population.** Draw  $n$  uniformly from  $\{5, 6, 7\}$ . Sample  $n - 1$  distinct centers without replacement from  $24\{-3, \dots, 3\}^2$  and set the final center to the negative sum. Reject unsupported final centers, repeated centers, overlapping disks, and quarter-turn symmetry. Each retained center carries a radius-eight disk; the rasterized disks have equal area. Their zero center sum therefore gives the same exact foreground centroid for all four quarter-turn orientations. Each initial plot is  $512 \times 512$ , stored as a  $512 \times 544$  RGB PNG including its label strip. The four labeled candidate panels form a  $1024 \times 1088$  montage. Both fit the explicit 1,344-pixel edge bound without downsampling. Unknown native provider processing is not covered by this byte-level statement.

Reference displacement, transport sign, reference task, candidate permutation and layout coefficient permutations are drawn before geometry acceptance. The fixed support, boundary, noncollinearity, geometry and pixel-oracle gates define the accepted population; no response is used to accept a proposal. Every proposed specification and rejection reason is retained. Formal draws exclude the fixed set of 64 calibration shape orbits and retain repetitions within the formal sample. Unlike the distinct-orbit calibration screen, these formal acceptance rules do not adapt to earlier formal draws. The frozen population contains 2,304 accepted groups from 15,172 proposals (12,868 rejections), with 941, 728 and 635 groups containing five, six and seven disks, respectively. It contains 2,195 distinct quarter-turn shape orbits; all 109 repeated orbit draws are retained.

**A noncollinear extension of the matched transport.** Keep the reference centroid  $c$  with coordinates in  $\{\pm 25, \pm 50\}$ . For the three wrong candidates use

$$b_i = e_i c + s k_i J c, \quad \{e_i\} = \{-1/5, 0, 1/5\}, \quad \{k_i\} = \{1/5, 2/5, 3/5\}, \quad s \in \{-1, +1\},$$

with independent permutations of the two coefficient sets. Reject any layout with a collinear triple of candidate centroids. Apply the same reference-fixed transport in Eq. (1), using cosine  $3/5$  and sines  $\pm 4/5$ . All translations are integer valued. The original wrong-candidate squared radius is  $(e_i^2 + k_i^2)\|c\|^2 < \|c\|^2$ . After transport it is

$$\frac{\|b_i(\theta)\|^2}{\|c\|^2} = \frac{4}{5} - \frac{4}{5}e_i + e_i^2 + k_i^2 \mp \frac{8}{5}k_i.$$

On the stated coefficient sets this is strictly below one for the sham and strictly above one for the intervention. The correct reference remains fixed while its radial rank reverses. Both transports preserve pairwise centroid distances, indexed-disk squared distances and equal foil movement magnitudes. They need not preserve silhouette similarity or perceptual difficulty. Noncollinearity removes the previous collinear-foil arrangement but does not establish absence of every relational shortcut. All frozen comparator outputs must still be reported.

**Input-only validation.** An independent verifier reads the retained RGB images and task prompts, extracts their connected foreground components, checks exact centroids and transport constraints, and solves all crossed questions by rotating and translating the observed foreground masks. It does not import the generator. A separate check canonicalizes the two actual initial-image orientations in each group to compare calibration and formal shape orbits. On the finalized inputs, the verifier

solves all 27,648 full-input question definitions (twelve per group), and the separate comparison finds zero orbit overlap between all 64 calibration and 2,304 formal groups. These checks establish input integrity; they are not measurements of a language model’s capability. Model request counts follow the separate cohort allocation below.

**Fixed cohorts and query allocation.** The generation protocol fixes 2,304 accepted groups. A PCG64 permutation with seed 202609172905 selects the shared 80-group complete core and the local configuration’s 2,048-group cohort; Sol uses all 2,304 groups. Every core group has twelve full-input questions (four per layout) and three candidate-only requests (one per layout). Every other group has only the matched reference sham/intervention pair. For a cohort of  $N$  groups and a core of  $C = 80$ , the planned request count is therefore  $2N + 13C$ : 5,648 for Sol and 5,136 for Qwen32. A candidate-only request is not counted four times simply because its balanced-quartet score can be derived analytically.

This allocation gives the two estimands different denominators. Crossed accuracy and complete-quartet success use the fixed 80-group core, while reference sensitivity uses the entire model-specific cohort. The local cohort overlaps Sol’s, so these are not independent model samples; the simultaneous-coverage union bound does not require cross-model independence. Call order is randomized before responses, and no completion determines the next scheduled question. Calibration, transport checks and pre-formal interface-calibration attempts remain outside the formal request totals.

## G DISK-FAMILY CALIBRATION AND FIXED CONFIGURATION SELECTION

All observations in this section are calibration and are excluded from the formal request counts and effect estimates. The disk sampler, renderer and acceptance rules were fixed before the disk-family screening calls. Model selection never used a sham–intervention effect.

**Interface feasibility.** We first attempted one preselected additional API configuration, Gemma-4-31B-IT, on 16 original-layout questions from the separate orientation calibration set. All HTTP responses returned, but only 2/16 contained a valid A–D JSON answer at an output cap of 2,048 tokens. A single declared interface repair raised the cap to 8,192 on the same 16 questions; 10/16 were valid, below the fixed 15/16 interface threshold. Truncated or missing answer content was retained. This establishes an interface failure under these settings, not the configuration’s spatial accuracy. We stopped that route and recorded one local fallback, Qwen2.5-VL-32B-Instruct, which returned 16/16 valid answers under its fixed greedy, bfloat16, batch-one configuration.

**Original-layout screening on new shapes.** The fixed calibration set contains 64 distinct disk-shape orbits and no formal test groups. Sol is evaluated on the first 16 groups and Qwen32 on all 64; each group contributes four original-layout questions. The predetermined rule requires at least 95% valid answers and  $\hat{a} - \sqrt{\log(40)/(2G)} > 1/2$ , where  $\hat{a}$  is the average four-question group score and  $G$  is the planned group count. Sol returns 64/64 valid answers and 57 correct, giving  $\hat{a} = 89.06\%$  and screening statistic 55.11%. Qwen32 returns 256/256 valid answers and 52 correct, giving  $\hat{a} = 20.31\%$  and statistic 3.34%. The former passes and the latter fails the unchanged operational screen. Because calibration dynamically excludes duplicate orbits, these screening statistics are not presented as independent-sample population confidence bounds. Formal crossed competence is assessed separately on the fixed 80-group core using the predeclared eight-test Holm family; its results appear in Section 4.3.

Before formal calls, we disclose Qwen32’s role as a low-accuracy comparison. It is not a second screen-positive configuration, and a narrow matched interval alone cannot establish spatial capability. The formal sampler excludes the fixed calibration orbit set but retains repeats within the formal draws. A separate pixel-based check compares actual initial-image orbits between calibration and formal populations.

**Concurrency and checkpoint checks.** An additional 64 calls repeat the Sol calibration inputs at eight concurrent requests to validate concurrent transport handling. All 64 return HTTP 200 and valid A–D answers, with no missing usage. Correctness is not used to choose concurrency

or revise the capability screen. Every local Qwen32 weight shard (18 files) and runtime configuration/tokenizer file (12 files) matches the official public checkpoint’s content objects at revision 7cfeb30d71a1f4f49a57592323337a4a4727301da.

## H RESOURCE AMENDMENT AND THE ORIGINAL STOPPED ANALYSIS

The disk study has one frozen planned cohort and two collection segments. The resource amendment was finalized after the original collection had stopped and its outcomes had been analyzed. It is therefore a post-launch, post-analysis amendment, not a new preregistration or an independent replication. Both the original-stop and resource-amended analyses are reported, irrespective of direction or precision. The earlier studies and the local Qwen32 response stream are reused unchanged.

**An unchanged task set.** The original Sol manifest records a clean budget stop after 4,793 of 5,648 planned requests, with every launched request recorded and every worker drained. Attempted identifiers are exactly the first 4,793 entries in the frozen dispatch order. The amendment admits only the remaining 855 task objects, copied byte for byte in that order. Selection uses attempted identifiers, dispatch order and resource accounting, not correctness. Previously attempted requests are excluded even if their answers were invalid, timed out or had unknown usage. There are no retries, replacement tasks or additional groups. The planned Sol reference cohort remains 2,304 groups and its crossed core remains 80 groups.

**Two actual collection records.** The continuation retains the requested model `gpt-5.6-sol`, low reasoning, an 8,192-output-token cap, a 180-second request timeout, eight workers, Pillow 10.2.0 and the frozen request implementation. Images, prompts, labels, seeds and task order remain fixed. The continuation has its own raw stream and genuine manifest; a combined original manifest is not synthesized. The original stopped on 18 September 2026 at 01:08 UTC, and the amendment was recorded that day at 09:14 UTC. The continuation ran from 09:19 to 09:42 UTC and recorded all 855 suffix requests; the gap between segments was 8 hours 11 minutes. The reported alias is `gpt-5.6-sol` on 4,759 original and 854 continuation responses; it is unreported on 34 and one, respectively. Shared requested or reported aliases do not establish an unchanged provider implementation; stable response kernels and conditional group independence remain assumptions.

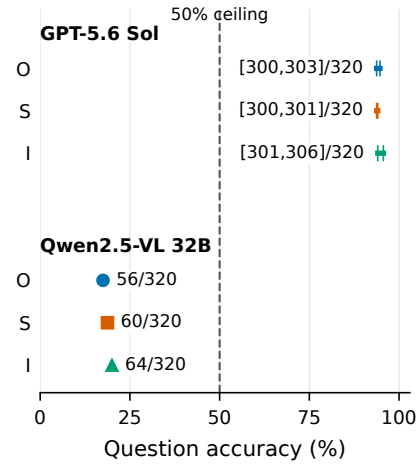
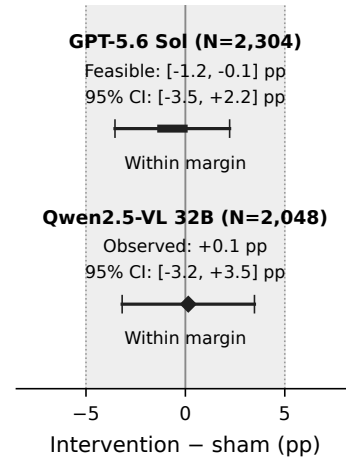
**Unchanged analysis and separate views.** The original scorers and statistical functions analyze the full original task lists. The eight-test Holm family still contains six crossed tests and two reference tests. The primary precision calculation retains the five equal-weight bets, four-tail allocation and strict  $\pm 5$ -point criterion in Appendix N; the eight-tail KL comparator remains separate. Completed invalid answers are errors. Attempted unknowns and any unattempted requests retain their adverse feasible completions. Planned denominators never shrink.

The original-stop view uses only the original streams, while the amended view additionally uses the distinct continuation stream. Shared responses are counted once in each view and are not independent replication evidence. The fixed- $N$  guarantee concerns each stated two-model analysis under its assumptions; it is not a claim of simultaneous coverage across successive reports or an optional-stopping confidence sequence.

**Replay.** Offline replay validates both segment manifests and regenerates `original-stop` and `resource-amended` analysis directories, with `comparison-receipt.json` recording their relationship. The scientific amendment binds the original freeze, finalized stream and manifest, continuation suffix and runtime/replay sources. The original frozen protocol, raw records and stopped outputs remain unchanged.

## I COMPLETE RESOURCE-AMENDED DISK-FAMILY RESULTS

The resource-amended view combines the two Sol segments described in Appendix H; Table 15 retains the original-stop view. Sol retains all 2,304 reference groups, Qwen32 all 2,048 prespecified groups, and each crossed core all 80 shared groups. O/S/I mean original/sham/intervention. Known

**Resource-amended disk cohort****a Shared crossed core****b Full-cohort reference effect**

O: original; S: sham; I: intervention. Full planned denominators retained.

a: descriptive accuracy; b: two-model simultaneous 95% fixed-N interval.

Sol adds the frozen unattempted suffix; Qwen is unchanged. Same cohort, not a replication.

Figure 10: **Crossed competence and average tolerance separate on disk tasks.** Left: full-plan accuracy bounds over 320 questions in each 80-group core; these are completion bounds, not confidence whiskers. Right: primary simultaneous 95% intervals on 2,304 Sol and 2,048 Qwen32 reference pairs. Sol’s 26 unknown pairs retain adverse completions. Both means meet the shaded  $\pm 5$ -point criterion despite their contrasting crossed results. The panels have different denominators.

invalid answers are observed failures; unavailable outcomes and unattempted tasks retain feasible completion bounds in the original denominators.

| Model / state                     | Recorded / planned | Unsent | Unknown pairs | W/L     | Effect (pp)    | 95% CI (pp)    |
|-----------------------------------|--------------------|--------|---------------|---------|----------------|----------------|
| GPT-5.6 Sol / Original stop       | 4793/5648          | 855    | 654           | 87/87   | [-16.2, +14.8] | [-19.6, +18.1] |
| GPT-5.6 Sol / Resource amended    | 5648/5648          | 0      | 26            | 111/126 | [-1.2, -0.1]   | [-3.5, +2.2]   |
| Qwen2.5-VL 32B / Original stop    | 5136/5136          | 0      | 0             | 200/197 | +0.1           | [-3.2, +3.5]   |
| Qwen2.5-VL 32B / Resource amended | 5136/5136          | 0      | 0             | 200/197 | +0.1           | [-3.2, +3.5]   |

The two states reuse the same frozen cohort; they are not independent replications. Qwen is unchanged. Unknown pairs include unavailable outcomes and unattempted tasks; each state retains its original full  $N$ . Intervals are simultaneous across two models within each state, without a joint guarantee across reporting states.

Table 15: Preserved original stop and resource-amended reporting on the same frozen cohort. W/L are known reference discordances. Feasible effects and confidence intervals retain all planned groups.

**Same-cohort reporting states.**

**Recorded requests.** The preserved Sol segment has 4,793 recorded requests; the continuation completes its 855/855 scheduled requests. Combined Sol records 5,648/5,648, and Qwen32 records 5,136/5,136. There are 10,784 actual disk calls: 10,739 valid, 10 completed invalid and 35 recorded unknown; none remain unattempted.

**All eight discovery tests.** The six crossed tests compare the expected four-question group score to  $1/2$ ; the two reference tests use the maximum exact McNemar  $p$ -value over all feasible unknown completions. These eight tests form one Holm family. The McNemar null requires independent

| Model  | Layout | Planned | Recorded | Valid | Invalid | Rec. unknown | Unattempted |
|--------|--------|---------|----------|-------|---------|--------------|-------------|
| Sol    | O      | 400     | 400      | 395   | 1       | 4            | 0           |
| Sol    | S      | 2,624   | 2,624    | 2,605 | 3       | 16           | 0           |
| Sol    | I      | 2,624   | 2,624    | 2,603 | 6       | 15           | 0           |
| Qwen32 | O      | 400     | 400      | 400   | 0       | 0            | 0           |
| Qwen32 | S      | 2,368   | 2,368    | 2,368 | 0       | 0            | 0           |
| Qwen32 | I      | 2,368   | 2,368    | 2,368 | 0       | 0            | 0           |

Table 16: Complete model–layout request accounting, including candidate-only calls. Planned requests are not presented as actual calls.

pairs with fair discordant directions; it is not the general heterogeneous average-zero null covered by the separate bounded-mean interval.

| Model  | Test          | Raw $p$                | Holm $p$               |
|--------|---------------|------------------------|------------------------|
| Sol    | O             | $5.01 \times 10^{-14}$ | $3.51 \times 10^{-13}$ |
| Sol    | S             | $5.01 \times 10^{-14}$ | $3.51 \times 10^{-13}$ |
| Sol    | I             | $3.23 \times 10^{-14}$ | $2.58 \times 10^{-13}$ |
| Sol    | Reference I–S | 0.9496                 | 1.0000                 |
| Qwen32 | O             | 1.0000                 | 1.0000                 |
| Qwen32 | S             | 1.0000                 | 1.0000                 |
| Qwen32 | I             | 1.0000                 | 1.0000                 |
| Qwen32 | Reference I–S | 0.9201                 | 1.0000                 |

Table 17: All prespecified discovery tests; no model or layout is selected after outcomes.

| Model  | Layout | Complete groups | Accuracy bounds (%) | $K = 4$ count bounds | 95% lower (%) |
|--------|--------|-----------------|---------------------|----------------------|---------------|
| Sol    | O      | 77/80           | [93.75, 94.69]      | [61, 64]             | 80.07         |
| Sol    | S      | 79/80           | [93.75, 94.06]      | [60, 61]             | 80.07         |
| Sol    | I      | 75/80           | [94.06, 95.62]      | [62, 67]             | 80.38         |
| Qwen32 | O      | 80/80           | [17.50, 17.50]      | [0, 0]               | 3.82          |
| Qwen32 | S      | 80/80           | [18.75, 18.75]      | [0, 0]               | 5.07          |
| Qwen32 | I      | 80/80           | [20.00, 20.00]      | [0, 0]               | 6.32          |

Table 18: Crossed accuracy uses all 320 planned core questions per layout. Correct-quartet bounds use all 80 planned groups. The one-sided accuracy bound is descriptive and unadjusted; discovery decisions use Table 17.

| Model  | Layout | $K = 0$ | $K = 1$ | $K = 2$ | $K = 3$ | $K = 4$ | Incomplete $K$ ranges:count |
|--------|--------|---------|---------|---------|---------|---------|-----------------------------|
| Sol    | O      | 0       | 0       | 1       | 15      | 61      | 3–4:3                       |
| Sol    | S      | 0       | 0       | 0       | 19      | 60      | 3–4:1                       |
| Sol    | I      | 0       | 0       | 1       | 12      | 62      | 3–4:5                       |
| Qwen32 | O      | 32      | 41      | 6       | 1       | 0       | –                           |
| Qwen32 | S      | 30      | 40      | 10      | 0       | 0       | –                           |
| Qwen32 | I      | 24      | 48      | 8       | 0       | 0       | –                           |

Table 19: Quartet correctness counts for complete groups, with every feasible range retained for incomplete groups. Incomplete groups are not removed from primary tests.

The primary interval targets average expected reference correctness difference on all frozen groups, conditional on the tasks, under stable response kernels and independent group-level differences. It is fixed- $N$ , not an optional-stopping confidence sequence. Each reporting state retains adverse missing completions;  $N$  does not shrink to completed pairs. The resource amendment follows the observed original stop. No optional-stopping or joint-across-states coverage guarantee is claimed. No narrow interval establishes model capability, renderer invariance or a sample-size-isolated effect.

| Model  | $N$   | Complete | $W$ | $L$ | Both correct | Both wrong | Unknown pairs |
|--------|-------|----------|-----|-----|--------------|------------|---------------|
| Sol    | 2,304 | 2,278    | 111 | 126 | 2,031        | 10         | 26            |
| Qwen32 | 2,048 | 2,048    | 200 | 197 | 256          | 1,395      | 0             |

Table 20: Reference paired accounting.  $W$  means intervention correct/sham wrong;  $L$  means the reverse. Correct/incorrect refers to the required full-input reference task.

| Model  | Feasible effect | Primary betting | Separated KL  | Complete-pair mean |
|--------|-----------------|-----------------|---------------|--------------------|
| Sol    | [-1.22, -0.09]  | [-3.54, 2.23]   | [-4.22, 2.92] | -0.66              |
| Qwen32 | [0.15, 0.15]    | [-3.18, 3.48]   | [-4.02, 4.31] | 0.15               |

Table 21: Reference intervention-minus-sham effects in percentage points. Feasible effects include all unknown completions; the complete-pair mean is descriptive only. Betting allocates .05/4 across four tails, while separated KL uses its own .05/8 eight-tail family; neither replaces the other.

| Model  | $D \in \{-1, 0\}$ | $D \in \{0, 1\}$ | $D \in \{-1, 0, 1\}$ |
|--------|-------------------|------------------|----------------------|
| Sol    | 13                | 13               | 0                    |
| Qwen32 | 0                 | 0                | 0                    |

Table 22: Unknown pair supports preserve whichever opposite-side correctness is known. Their sum equals the unknown-pair count in Table 20; a partially known pair is not automatically assigned the full range.

| Model  | Layout | All four    | Reference | Other three |
|--------|--------|-------------|-----------|-------------|
| Sol    | O      | 300/320; ?3 | 72/80; ?2 | 228/240; ?1 |
| Sol    | S      | 300/320; ?1 | 77/80; ?0 | 223/240; ?1 |
| Sol    | I      | 301/320; ?5 | 76/80; ?1 | 225/240; ?4 |
| Qwen32 | O      | 56/320; ?0  | 17/80; ?0 | 39/240; ?0  |
| Qwen32 | S      | 60/320; ?0  | 17/80; ?0 | 43/240; ?0  |
| Qwen32 | I      | 64/320; ?0  | 15/80; ?0 | 49/240; ?0  |

Table 23: Fixed-core masks: each cell is known correct/planned questions; ? gives unknown questions. These descriptive masks are assigned before model responses and do not isolate ordinary one-factor ablations.

| Model | Contrast | Mask              | Complete | Feasible effect | Bootstrap 95%  |
|-------|----------|-------------------|----------|-----------------|----------------|
| Sol   | I-S      | All four          | 74/80    | [0.00, 1.88]    | [-1.35, 5.41]  |
| Sol   | I-S      | Reference         | 79/80    | [-1.25, 0.00]   | [-6.33, 6.33]  |
| Sol   | I-S      | Other three       | 75/80    | [0.42, 2.50]    | [-1.33, 6.22]  |
| Sol   | I-S      | Ref. minus others | 74/80    | [-3.75, -0.42]  | [-10.36, 4.95] |
| Sol   | S-O      | All four          | 76/80    | [-0.94, 0.31]   | [-3.95, 2.63]  |
| Sol   | S-O      | Reference         | 78/80    | [3.75, 6.25]    | [-3.85, 11.54] |
| Sol   | S-O      | Other three       | 78/80    | [-2.50, -1.67]  | [-5.98, 1.28]  |
| Sol   | S-O      | Ref. minus others | 76/80    | [5.42, 8.75]    | [-2.63, 14.91] |
| Sol   | I-O      | All four          | 72/80    | [-0.62, 1.88]   | [-2.08, 4.86]  |
| Sol   | I-O      | Reference         | 77/80    | [2.50, 6.25]    | [-2.60, 10.39] |
| Sol   | I-O      | Other three       | 75/80    | [-1.67, 0.42]   | [-3.11, 4.00]  |
| Sol   | I-O      | Ref. minus others | 72/80    | [2.08, 7.92]    | [-3.70, 11.57] |

Table 24: All fixed-core mask contrasts for Sol. Effects are percentage points; bootstrap intervals use complete groups only and are descriptive/unadjusted. Feasible bounds retain all 80 groups.

| Model  | Contrast | Mask              | Complete | Feasible effect | Bootstrap 95%  |
|--------|----------|-------------------|----------|-----------------|----------------|
| Qwen32 | I-S      | All four          | 80/80    | [1.25, 1.25]    | [-3.44, 5.94]  |
| Qwen32 | I-S      | Reference         | 80/80    | [-2.50, -2.50]  | [-11.25, 6.25] |
| Qwen32 | I-S      | Other three       | 80/80    | [2.50, 2.50]    | [-2.92, 7.92]  |
| Qwen32 | I-S      | Ref. minus others | 80/80    | [-5.00, -5.00]  | [-15.00, 5.00] |
| Qwen32 | S-O      | All four          | 80/80    | [1.25, 1.25]    | [-3.12, 5.62]  |
| Qwen32 | S-O      | Reference         | 80/80    | [0.00, 0.00]    | [-7.50, 7.50]  |
| Qwen32 | S-O      | Other three       | 80/80    | [1.67, 1.67]    | [-4.17, 7.08]  |
| Qwen32 | S-O      | Ref. minus others | 80/80    | [-1.67, -1.67]  | [-10.83, 7.92] |
| Qwen32 | I-O      | All four          | 80/80    | [2.50, 2.50]    | [-1.56, 6.25]  |
| Qwen32 | I-O      | Reference         | 80/80    | [-2.50, -2.50]  | [-10.00, 3.75] |
| Qwen32 | I-O      | Other three       | 80/80    | [4.17, 4.17]    | [-0.42, 8.33]  |
| Qwen32 | I-O      | Ref. minus others | 80/80    | [-6.67, -6.67]  | [-14.17, 1.25] |

Table 25: All fixed-core mask contrasts for Qwen32. Effects are percentage points; bootstrap intervals use complete groups only and are descriptive/unadjusted. Feasible bounds retain all 80 groups.

| Model  | Layout | Planned | Recorded | Valid | Unknown<br>incl. missing | Quartet<br>bounds (%) | Ref.<br>matches |
|--------|--------|---------|----------|-------|--------------------------|-----------------------|-----------------|
| Sol    | O      | 80      | 80       | 79    | 1                        | [24.69, 25.00]        | 34              |
| Sol    | S      | 80      | 80       | 80    | 0                        | [25.00, 25.00]        | 28              |
| Sol    | I      | 80      | 80       | 80    | 0                        | [25.00, 25.00]        | 31              |
| Qwen32 | O      | 80      | 80       | 80    | 0                        | [25.00, 25.00]        | 15              |
| Qwen32 | S      | 80      | 80       | 80    | 0                        | [25.00, 25.00]        | 19              |
| Qwen32 | I      | 80      | 80       | 80    | 0                        | [25.00, 25.00]        | 12              |

Table 26: Candidate-only accounting: one request is scored analytically across four balanced labels and is never counted as four calls. There is no unique-answer accuracy for this input. Reference matches describe response preference only.

**All frozen comparator predictions.** These are CPU input-side predictions, not additional model requests. The source oracle uses generation metadata; the pixel-mask oracle uses foreground geometry. Radial/outlier rules and the centroid executor are distinct restrictions. Perfect geometric solvability does not measure learned-model capability. Submitted-resolution predictions refer to the explicit RGB/PNG transform, not unknown native model processing.

| Population     | Comparator         | O             | S                   | I                   |
|----------------|--------------------|---------------|---------------------|---------------------|
| Core full      | Radial, source     | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Outlier, source    | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Radial, submitted  | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Outlier, submitted | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Centroid executor  | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Source oracle      | 320/320; a320 | 320/320; a320       | 320/320; a320       |
| Core full      | Pixel-mask oracle  | 320/320; a320 | 320/320; a320       | 320/320; a320       |
| Reference      | Radial, source     | 80/80; a80    | 2,304/2,304; a2,304 | 0/2,304; a2,304     |
| Reference      | Outlier, source    | 80/80; a80    | 2,304/2,304; a2,304 | 2,304/2,304; a2,304 |
| Reference      | Radial, submitted  | 80/80; a80    | 2,304/2,304; a2,304 | 0/2,304; a2,304     |
| Reference      | Outlier, submitted | 80/80; a80    | 2,304/2,304; a2,304 | 2,304/2,304; a2,304 |
| Reference      | Centroid executor  | 0/80; a80     | 0/2,304; a2,304     | 2,304/2,304; a2,304 |
| Reference      | Source oracle      | 80/80; a80    | 2,304/2,304; a2,304 | 2,304/2,304; a2,304 |
| Reference      | Pixel-mask oracle  | 80/80; a80    | 2,304/2,304; a2,304 | 2,304/2,304; a2,304 |
| Candidate-only | Radial, source     | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Outlier, source    | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Radial, submitted  | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Outlier, submitted | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Centroid executor  | 0/80; –       | 0/80; –             | 0/80; –             |
| Candidate-only | Source oracle      | 0/80; –       | 0/80; –             | 0/80; –             |
| Candidate-only | Pixel-mask oracle  | 0/80; –       | 0/80; –             | 0/80; –             |

Table 27: All frozen comparator cells for Sol. Full-input cells show correct/planned; a gives available predictions. Candidate-only cells show available/planned and analytic balanced-quartet score when predictions are available. Dashes mark unavailable predictions, not uniquely scored wrong answers.

| Population     | Comparator         | O             | S                   | I                   |
|----------------|--------------------|---------------|---------------------|---------------------|
| Core full      | Radial, source     | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Outlier, source    | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Radial, submitted  | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Outlier, submitted | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Centroid executor  | 80/320; a320  | 80/320; a320        | 80/320; a320        |
| Core full      | Source oracle      | 320/320; a320 | 320/320; a320       | 320/320; a320       |
| Core full      | Pixel-mask oracle  | 320/320; a320 | 320/320; a320       | 320/320; a320       |
| Reference      | Radial, source     | 80/80; a80    | 2,048/2,048; a2,048 | 0/2,048; a2,048     |
| Reference      | Outlier, source    | 80/80; a80    | 2,048/2,048; a2,048 | 2,048/2,048; a2,048 |
| Reference      | Radial, submitted  | 80/80; a80    | 2,048/2,048; a2,048 | 0/2,048; a2,048     |
| Reference      | Outlier, submitted | 80/80; a80    | 2,048/2,048; a2,048 | 2,048/2,048; a2,048 |
| Reference      | Centroid executor  | 0/80; a80     | 0/2,048; a2,048     | 2,048/2,048; a2,048 |
| Reference      | Source oracle      | 80/80; a80    | 2,048/2,048; a2,048 | 2,048/2,048; a2,048 |
| Reference      | Pixel-mask oracle  | 80/80; a80    | 2,048/2,048; a2,048 | 2,048/2,048; a2,048 |
| Candidate-only | Radial, source     | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Outlier, source    | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Radial, submitted  | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Outlier, submitted | 80/80; 25.00% | 80/80; 25.00%       | 80/80; 25.00%       |
| Candidate-only | Centroid executor  | 0/80; –       | 0/80; –             | 0/80; –             |
| Candidate-only | Source oracle      | 0/80; –       | 0/80; –             | 0/80; –             |
| Candidate-only | Pixel-mask oracle  | 0/80; –       | 0/80; –             | 0/80; –             |

Table 28: All frozen comparator cells for Qwen32. Full-input cells show correct/planned; a gives available predictions. Candidate-only cells show available/planned and analytic balanced-quartet score when predictions are available. Dashes mark unavailable predictions, not uniquely scored wrong answers.

## J POST-HOC SENSITIVITY TO THE PRACTICAL MARGIN

This display is a deterministic readout of the unchanged resource-amended confidence intervals, added after the model results. It does not recompute intervals, alter any response or choose a new primary practical criterion. For a fixed interval  $[L, U]$ , strict containment in  $(-\varepsilon, \varepsilon)$  is equivalent to  $\varepsilon > \varepsilon_* := \max(-L, U)$ . Because containment is strict,  $\varepsilon_*$  is an infimum threshold, not an attained minimum margin; equality does not pass. Decisions below use unrounded endpoints.

The primary two-model simultaneous intervals and the separate prespecified KL comparator retain their original definitions and tail allocations. Reading several margins from one fixed interval introduces no new fitted intervals; it is a post-hoc precision summary, not a new confirmatory test or a utility-calibrated threshold. The original 5-point decision is unchanged.

| Model  | Interval family | $\varepsilon_*$ (pp) | 3 pp | 3.5 pp | 4 pp | 5 pp |
|--------|-----------------|----------------------|------|--------|------|------|
| Sol    | Primary         | 3.5413               | No   | No     | Yes  | Yes  |
| Sol    | Separated KL    | 4.2159               | No   | No     | No   | Yes  |
| Qwen32 | Primary         | 3.4751               | No   | Yes    | Yes  | Yes  |
| Qwen32 | Separated KL    | 4.3107               | No   | No     | No   | Yes  |

Table 29: Strict within-margin readings of the fixed resource-amended intervals. The primary and KL procedures are shown separately; the comparator is not selected in place of the primary decision.

## K REPEATED RESPONSES AND CATEGORICAL LAYOUT DEPENDENCE

The separately frozen Qwen32 and Sol experiments use fresh responses to the same fixed subset of existing bitmap and disk inputs, not independent datasets. Analysis conditions on those inputs. Their response-independence assumptions are stronger than the arbitrary within-quartet dependence allowed by the crossed omission tests.

**Observation and estimand.** For each fixed source group, collect four stateless responses in randomized global call order: two to sham and two to intervention. Candidate letters are never permuted between replicas. The categorical alphabet consists of A–D and a completed-invalid category  $\perp$ . Infrastructure failures and requests never launched are unknown, not  $\perp$ . Under a stable response kernel and independent repeat randomness,

$$\mathbb{E}\mathbf{1}\{S_0 = S_1\} = \sum_k p_{g,S}(k)^2, \quad \mathbb{E}\mathbf{1}\{S_a = I_b\} = \sum_k p_{g,S}(k)p_{g,I}(k).$$

Substitution in Equation 4 gives  $\mathbb{E}T_g = \frac{1}{2} \sum_k [p_{g,S}(k) - p_{g,I}(k)]^2$ . Thus the equally weighted group mean estimates a nonnegative parameter in  $[0, 1]$ . Its sample estimate can be negative and is retained without truncation. Two deterministic but different answers yield  $T_g = 1$ , even if both are wrong; identical uniform kernels have zero expected statistic despite frequent raw answer disagreements. The category-specific summands are retained descriptively so an effect due to formatting cannot be called exclusively a spatial-choice shift.

**Finite range and missing outcomes.** If both within-layout pairs agree,  $T_g$  is zero or one. If exactly one agrees, the cross-layout equality count is at most two and  $T_g$  is zero or one half. If neither agrees, that count is at most two and  $T_g$  is between minus one half and zero. The exact range is therefore  $[-1/2, 1]$ . Each unknown response has five feasible categorical completions. Enumerating at most  $5^4$  combinations gives exact feasible bounds  $[T_g^-, T_g^+]$ . Their means retain all planned groups and require no missing-at-random assumption.

**Prespecified discovery family.** The null  $\eta = 0$  implies  $p_{g,S} = p_{g,I}$  for every group because every  $\eta_g$  is nonnegative. For independent group response draws, Hoeffding’s inequality and  $T_g^- \leq T_g$  give the valid conservative value

$$p_H = \exp\{-8N[\max(0, \overline{T}^-)]^2/9\}.$$

When every planned response is observed, the null also makes each group’s four responses exchangeable. Conditional on their unordered pool, each of the six index assignments to two sham and two intervention positions is equally likely. This remains true when categorical values repeat. The exact upper tail includes ties. We compute the six possible integer scores  $4T_g$  and convolve their integer count distributions across groups, dividing the tail count by  $6^N$ . Independent groups justify this product assignment law; equal marginals alone without exchangeable independent repeats do not.

If any response in a family is unavailable, we set that family’s permutation component to one. This is pointwise no smaller than the latent complete-data permutation value and hence remains conservative under arbitrary missingness. We do not drop incomplete groups or insert guessed values into the observed permutation pool. The frozen combined test is  $p_{\text{comb}} = \min(1, 2 \min(p_H, p_{\text{perm}}))$ . The union bound gives  $\Pr(p_{\text{comb}} \leq \alpha) \leq \Pr(p_H \leq \alpha/2) + \Pr(p_{\text{perm}} \leq \alpha/2) \leq \alpha$ . Holm correction then controls the two-test family comprising bitmap and disk within each predeclared model, without a joint across-model guarantee. Neither analysis selects the more favorable primitive after seeing its result.

**Separate effect precision.** For two simultaneous two-sided intervals, allocate 0.0125 to each tail and set  $r_N = (3/2)\sqrt{\log(1/0.0125)/(2N)}$ . Intersect  $[\bar{T}^- - r_N, \bar{T}^+ + r_N]$  with  $[0, 1]$  and explicitly retain an empty intersection if one occurs. The no-missingness radii are 19.63 points for  $N = 128$  and 13.88 for  $N = 256$ ; exact discovery can reject while these conservative intervals include zero. Discovery and precision are separate error-control families, without a joint 95% claim over both.

**Scope of the response assumptions.** Inference assumes stable categorical response kernels, independent repeats and independent group response draws, conditional on frozen inputs and the generation contract. A timeout does not establish server completion or cancellation. For local greedy inference, agreement of repeated outputs provides an observed control, not a guarantee of bitwise reproducibility under all hardware, batching or software changes. These experiments identify neither an internal model mechanism nor downstream agent reliability.

## L EXACT CATEGORICAL PERMUTATION INFERENCE WITH UNAVAILABLE RESPONSES

The original bitmap/disk repeat analysis sets its permutation branch to one if any planned response is unknown. Its secondary extension was fixed after administrative missing counts were known but before the current Sol category or effect records were inspected. It leaves the original primary tests and intervals unchanged. The later balanced-repeat study prospectively uses this exact test in its primary three-test family (Appendix D.1). Its mathematical validity does not require missing at random, but still requires well-defined latent responses for every planned slot and the independent stable-kernel null of Section 3.4. MMD-Miss bounds the unbiased squared MMD statistic with a Laplacian kernel and upper-bounds Monte Carlo permutation  $p$ -values without assumptions on the missingness mechanism (Zeng et al., 2025). The categorical  $2 + 2$  structure below instead yields an attained maximum through five ordered states; neither MMD nor worst-completion validity is claimed as new.

For a complete group, only response multiplicities  $2 + 2$  and  $2 + 1 + 1$  yield nonconstant permutation scores. Call them types  $A$  and  $B$ . Under the six equally likely assignments of two responses to each layout,  $4T$  equals 4,  $-2$  with probabilities  $1/3, 2/3$  for type  $A$ , and 2,  $-1$  with the same probabilities for type  $B$ . All other multiplicities give zero. Write  $A_+, A_-, B_+, B_-, 0$  for these five observed states. If there are  $n_A, n_B$  active groups and  $k = 2\#A_+ + \#B_+$ , the exact upper-tail probability is

$$p = \Pr(2U_A + U_B \geq k), \quad U_A \sim \text{Bin}(n_A, 1/3), \quad U_B \sim \text{Bin}(n_B, 1/3), \quad (13)$$

with independent binomials. Integer coefficient tails of  $(2+x^2)^{n_A}(2+x)^{n_B}$  give its exact numerator and denominator.

**Exact worst completion.** For  $V = 4T_{\text{perm}} - 4T_{\text{observed}}$ , ordered equal-weight three-point representations are

$$A_- : (0, 0, 6), \quad B_- : (0, 0, 3), \quad 0 : (0, 0, 0), \quad B_+ : (-3, -3, 0), \quad A_+ : (-6, -6, 0).$$

A common-uniform coupling gives the pointwise order  $A_- \succeq B_- \succeq 0 \succeq B_+ \succeq A_+$ . For each incomplete group, enumerate its at most  $5^4$  categorical completions and select the greatest attainable state in this order. Independent couplings show that the resulting sum  $V$  stochastically dominates every other global completion; it therefore attains the global maximum of  $\Pr(\sum_g V_g \geq 0)$ . Equation (13) computes that maximum. A representative completion witnesses the bound and is never treated as an observed or imputed answer.

If  $X$  is the full latent response vector and  $O$  its partially observed version,  $p_{\max}(O) \geq p(X)$  pointwise. Hence  $\Pr_0(p_{\max} \leq \alpha) \leq \Pr_0(p(X) \leq \alpha) \leq \alpha$  for arbitrary missingness. This is unconditional validity under the complete-data null, not validity conditional on an outcome-dependent missingness pattern, an anytime test, or a guarantee under dependent groups. Completed invalid responses are the fifth observed category; unavailable responses are not a sixth category. Ties are included, and positive rational probabilities are rounded upward for floating-point reporting.

**Checks and observed outcome.** Independent exact checks cover all 625 complete tuples, 1,296 partial tuples, 390,625 complete two-group configurations and 1,679,616 partial two-group configurations against 100 million compatible completions. Sol’s bitmap cohort has one selected  $A_+$  and 127 inactive states, giving  $p = 1/3$ ; its disk cohort has six  $A_-$  and 250 inactive states, giving  $p = 1$ . The separate audit family consists of the two original combined tests and the two worst-completion tests, with Holm correction across these four tests within a model. All four Sol adjusted values equal one. A retrospective application to the previously observed complete Qwen32 records reproduces the exact permutation branches and retains both discoveries in this larger family. No joint guarantee across models or with effect intervals is claimed.

#### PRIOR DIAGNOSTIC COMPONENTS

3DSRBench adopts CircularEval to assess answer-order sensitivity (Ma et al., 2025); CircularEval is a component inherited from prior work, not a procedure originated by 3DSRBench. Our discussion attributes the use of that component to 3DSRBench and makes no priority claim for it.

MM-R<sup>3</sup> already evaluates four repetitions of identical image–question inputs and compares their consistency with perturbed inputs (Chou et al., 2025). Repeated-input controls and comparisons against sampling variability are therefore prior art. Our repeated-response experiment applies the classical categorical MMD identity to the specific fixed-reference geometric contrast, with frozen inference and unknown-output rules; it does not claim to originate noise-controlled consistency measurement.

| Family | Measure       | $S_0$ | $S_1$ | $I_0$ | $I_1$ |
|--------|---------------|-------|-------|-------|-------|
| Bitmap | Correct / 128 | 47    | 51    | 39    | 40    |
| Disk   | Correct / 256 | 60    | 59    | 61    | 62    |

Table 30: Qwen32 correct answers in each repeat slot; all 1,536 outputs are valid. These are descriptive accuracy counts, not new equivalence decisions.

## M A POST-HOC THIRD-MOMENT PHASE BASELINE

The baseline tests whether a compact orientation feature can solve the crossed equal-centroid tasks. It is fully post hoc. Before its protocol was recorded, an exploratory initial-image census found zero third moment in 0/128 bitmap groups and 3/2,304 disk groups. Those groups remain in every relevant denominator. The later implementation freeze records execution choices; it does not make the experiment prospective. The classical moment identity below is an established construction (Flusser, 2000), not an algorithmic novelty claim.

**A sufficient descriptor for nonzero quarter-turn instances.** For foreground pixels use  $p_j = x_j - iy_j$ , since image rows increase downward, and let  $z_j = p_j - \bar{p}$ . The mean third complex central

moment is

$$m_3(X) = \frac{1}{N} \sum_{j=1}^N z_j^3. \quad (14)$$

Translation disappears under centering. Under a positive isotropic scale  $s$  and counterclockwise point rotation  $\theta$ , centered coordinates become  $se^{i\theta}z_j$ , so

$$m_3(sR_\theta X + t) = s^3 e^{3i\theta} m_3(X). \quad (15)$$

Thus clockwise and counterclockwise quarter turns multiply the moment by  $+i$  and  $-i$ , respectively. If  $m_3(X) \neq 0$ , the four quarter-turn versions have distinct unit phases  $u, -iu, -u, iu$ . Transforming the initial phase by the requested direction therefore selects a unique candidate in this ideal four-orientation problem. The label-ranking descriptor has two real components. It uses information beyond the initial centroid and is compatible with the omission bounds in the main text.

The nonzero assumption matters: an asymmetric shape can have zero third moment. Unrelated shapes may collide, and arbitrary-angle phase has a  $120^\circ$  ambiguity. These facts prevent a general shape-matching claim. Positive scaling preserves ideal phase, whereas raster interpolation, clipping and foreground-threshold changes need not preserve it.

**Fixed pixel interface and selective prediction.** The predictor accepts only initial RGB pixels, candidate RGB pixels and the public direction instruction. It removes the known 32-source-pixel label strips, splits the four panels in row-major order and applies the fixed blue thresholds  $B - R \geq 75$  and  $B - G \geq 37$ . Source framing is  $544 \times 512$  for an initial image and  $1088 \times 1024$  for candidates (height by width). Recorded transformed dimensions are separately validated; crop boundaries use the same fixed half-up scaling rule. These color and framing assumptions are specific to this renderer. No generation labels, filenames, shape identities or stored oracle predictions enter the predictor.

For  $S_k = \sum_j p_j^k$ , integer arithmetic gives

$$C_3 = N^2 S_3 - 3N S_1 S_2 + 2S_1^3 = N^3 m_3, \quad R_2 = N \sum_j |p_j|^2 - |S_1|^2. \quad (16)$$

The phase is  $C_3/|C_3|$ . Foreground count enters central-moment extraction; the second radial moment is additionally used only for the dimensionless abstention normalization:

$$\rho_3 = \frac{|m_3|}{(N^{-1} \sum_j |z_j|^2)^{3/2}} = \frac{|C_3|}{R_2^{3/2}}. \quad (17)$$

It abstains if  $R_2 \leq 0$  or  $\rho_3 \leq 10^{-12}$  in the initial image or *any* candidate. The latter gate is evaluated exactly as  $10^{24}|C_3|^2 \leq R_2^3$ . Otherwise it transforms the initial phase and chooses the nearest candidate by squared chord distance. A second distance within  $10^{-12}$  of the minimum causes whole-question abstention. There is no fallback to another moment, bitmap rotation, centroid matching, candidate deletion or task-specific exception. Empty foreground, invalid framing or instruction, degeneracy and ties retain distinct reason codes.

**Scope of the sufficiency argument.** If initial and candidate phase errors are bounded by  $a$  and  $b$ , respectively, put  $q = a + b$ . The correct phase is at distance at most  $q$  and a wrong phase at least  $\pi/2 - q$ . Thus  $q < \pi/4$  guarantees the correct ideal ordering. For  $0 \leq q < \pi/4$  and the implemented squared-chord tie tolerance  $\tau$ , the additional condition  $2(\cos q - \sin q) > \tau$  guarantees a unique selection before final floating-point rounding. Passing the near-zero gate alone supplies no such perturbation bound. The empirical census, rather than the covariance identity alone, establishes performance on the stored pixels.

**Populations and pixel-stage separation.** The source census comprises 29,184 full-input task definitions: 1,536 bitmap tasks from 128 groups and 27,648 disk tasks from 2,304 groups, with four questions in each of three layouts per group. These are input-side tasks, not additional model calls. Candidate-only rows have no unique target label and are accounted for separately: 384 bitmap and 6,912 disk rows. The requested Azure preparation stage contains 6,944 full-input tasks (1,536

bitmap and 5,408 disk); the local stage contains 6,432 (1,536 bitmap and 4,896 disk). They reuse source tasks rather than adding independent populations.

Submitted-stage claims require exact retained pixels or reconstruction with the recorded Pillow version and matching historical image hashes. A missing stage stays unavailable over its full planned denominator; it is never filled by a source-PNG prediction. This validation concerns explicit input preparation, not network delivery or unknown provider-native image processing. Pixel or prediction identity across stages is an equality check, not an independent robustness experiment. Predictions are committed before the separate evaluator joins the existing answer key; that procedural separation does not undo the experiment’s post-hoc status.

| Family | Pixel stage    | Planned | Correct | Wrong | Abst./unavail.     |
|--------|----------------|---------|---------|-------|--------------------|
| Tile   | Source PNG     | 1,536   | 1,536   | 0     | 0                  |
| Tile   | Azure explicit | 1,536   | 1,536   | 0     | 0                  |
| Tile   | Local explicit | 1,536   | –       | –     | 1,536 <sup>†</sup> |
| Disk   | Source PNG     | 27,648  | 27,612  | 0     | 36                 |
| Disk   | Azure explicit | 5,408   | 5,402   | 0     | 6                  |
| Disk   | Local explicit | 4,896   | –       | –     | 4,896 <sup>†</sup> |

Fully posthoc input-side solver, with pre-plan initial-moment exploration disclosed. Source rows cover all generated full-input tasks; explicit-stage rows reuse frozen requested subsets. <sup>†</sup>Recorded Pillow unavailable: these are unobserved stages, not measured zero accuracy. All solver abstentions retain their planned denominators.

Table 31: **Full-denominator third-moment phase census.** Source and requested pixel-stage populations are distinct and overlap. Solver abstentions and unavailable preparation stages remain explicit; no new model response is collected.

**Observed census and retained failures.** Table 31 gives every family/stage denominator. The source bitmap predictions are all correct. Disk correct-over-all-planned accuracy and coverage are both  $27,612/27,648 = 99.87\%$ ; conditional accuracy among returned predictions is 100%. All 36 source abstentions have code `initial_feature_near_zero`, in the three previously inspected groups 01583, 01866 and 01981. There are no wrong predictions or phase-distance ties. The all-four-correct count is 384/384 bitmap group/layout quartets and lies in  $[6,903, 6,912]$  out of 6,912 disk quartets when unresolved abstentions retain  $[0, 1]$  support. These are fixed-input census and completion counts, not sampling-confidence intervals.

The 6,944 requested Azure-stage outcomes match their source counterparts: 1,536 correct bitmap predictions and 5,402 correct plus six abstaining disk predictions. All 7,712 verified stage-image entries (640 bitmap and 7,072 disk) have RGB arrays identical to their source images. The local stage remains unavailable for all 6,432 planned tasks because its recorded Pillow 12.3.0 runtime is unavailable; the present 10.2.0 installation is not substituted. No local-stage baseline accuracy was observed. Neither this pixel equality nor the source census supplies independent model replication, new-renderer transfer, repeated-response evidence or a construction with changed reference-outlier identity.

## N FIXED-SAMPLE PAIRED PRECISION FOR THE CONSTELLATION PROTOCOL

Betting constructions for bounded-mean inference are established statistical tools (Waudby-Smith & Ramdas, 2024). The finite grid and equal weights below are implementation choices for this protocol, not a new statistical method. We give the fixed-sample argument explicitly because the target permits different group means. This argument does not provide a confidence sequence.

**Target and assumptions.** Condition on the information  $\mathcal{F}_0$  fixed before the original formal responses: the task set, model configuration, group count  $N \geq 1$ , and analysis plan. For reference group  $g$ , let  $Y_{g,I}, Y_{g,S} \in \{0, 1\}$  denote correctness under intervention and sham, and set

$$D_g = Y_{g,I} - Y_{g,S} \in \{-1, 0, 1\}, \quad \mu_g = \mathbb{E}[D_g \mid \mathcal{F}_0], \quad \Delta = \frac{1}{N} \sum_{g=1}^N \mu_g.$$

We assume the  $D_g$  are independent conditional on  $\mathcal{F}_0$ ; their distributions and means may differ. Dependence between the two responses within a group is unrestricted. The target is the average expected difference in reference correctness for the frozen groups. Extending it to another task population requires an additional sampling argument. Stateless requests and randomized order do not by themselves establish conditional independence.

**A fixed mixture of complete products.** For  $m \in [-1, 1]$  and fixed  $0 < \lambda \leq 1/2$ , define

$$E_\lambda^+(m) = \prod_{g=1}^N [1 + \lambda(D_g - m)], \quad E_\lambda^-(m) = \prod_{g=1}^N [1 + \lambda(m - D_g)]. \quad (18)$$

Use the same  $\lambda$  for every group within each product. The protocol fixes  $\Lambda = \{.1, .2, .3, .4, .5\}$  and  $w_\lambda = 1/5$ , and computes  $E^\pm(m) = \sum_{\lambda \in \Lambda} w_\lambda E_\lambda^\pm(m)$ . Thus the mixture is taken over complete products. Neither the grid nor its weights is selected using the formal responses. The resource amendment does not redefine  $\mathcal{F}_0$ , the planned  $N$  or this mixture.

**Fixed-sample validity under heterogeneous means.** All factors in Eq. (18) are nonnegative, including the possible zero factors at  $\lambda = 1/2$ . Under  $\Delta \leq m$ , conditional independence and the arithmetic–geometric mean inequality give

$$\mathbb{E}[E_\lambda^+(m) \mid \mathcal{F}_0] = \prod_{g=1}^N [1 + \lambda(\mu_g - m)] \quad (19)$$

$$\leq \left[ \frac{1}{N} \sum_{g=1}^N [1 + \lambda(\mu_g - m)] \right]^N = [1 + \lambda(\Delta - m)]^N \leq 1. \quad (20)$$

Consequently  $\mathbb{E}[E^+(m) \mid \mathcal{F}_0] \leq 1$ . Markov’s inequality bounds the probability of  $E^+(m) \geq 1/\alpha$  by  $\alpha$ . Applying the same argument to  $-D_g$  yields validity of  $E^-(m)$  under  $\Delta \geq m$ . No common mean, per-group null  $\mu_g \leq m$ , or identical distribution assumption is needed.

**Inversion and simultaneous coverage.**  $E^+(m)$  is nonincreasing in  $m$ , and  $E^-(m)$  is nondecreasing. Define closed confidence endpoints by

$$L(D) = \sup(\{-1\} \cup \{m \in [-1, 1] : E^+(m) \geq 1/\alpha\}), \quad (21)$$

$$U(D) = \inf(\{1\} \cup \{m \in [-1, 1] : E^-(m) \geq 1/\alpha\}). \quad (22)$$

The endpoint convention includes the boundary of the nonrejected set. At the empirical mean  $\bar{D}$ , the arithmetic–geometric mean inequality applied directly to each observed product gives  $E^+(\bar{D}), E^-(\bar{D}) \leq 1$ ; hence the interval is nonempty and contains  $\bar{D}$ . Monotonicity implies  $\Pr\{L(D) > \Delta \mid \mathcal{F}_0\} \leq \alpha$  and  $\Pr\{U(D) < \Delta \mid \mathcal{F}_0\} \leq \alpha$ . These bounds evaluate the statistic at the true  $\Delta$  and require no correction over candidate values of  $m$ .

The precision family reserves two model intervals and four one-sided tails, with  $\alpha = .05/4 = .0125$  and threshold  $1/\alpha = 80$  for every tail. Bonferroni therefore gives simultaneous coverage at least .95. The same allocation is retained if only one model configuration is ultimately reported. No independence between models is required. This is a separate family from the discovery tests and from previous experiments.

**Unknown outcomes and sharp feasible completions.** Retain every planned group. For an observed pair  $(a_g, b_g)$ , where either entry may be unknown, let

$$\mathcal{D}_g = \{x - y : x \in \mathcal{Y}(a_g), y \in \mathcal{Y}(b_g)\}, \quad \mathcal{Y}(z) = \begin{cases} \{z\}, & z \in \{0, 1\}, \\ \{0, 1\}, & z \text{ is unknown.} \end{cases}$$

For example,  $(1, \text{unknown})$  permits  $\{0, 1\}$ , whereas  $(0, \text{unknown})$  permits  $\{-1, 0\}$ . A both-unknown pair permits  $\{-1, 0, 1\}$ . A completed invalid answer is an observed zero correctness, not an unknown response. Set  $D_g^- = \min \mathcal{D}_g$  and  $D_g^+ = \max \mathcal{D}_g$ , and report

$$[L(D^-), U(D^+)]. \quad (23)$$

The endpoints above are coordinatewise nondecreasing in the observations. Thus this interval contains the complete-data interval for every feasible completion. Its lower and upper endpoints are each attained by a feasible completion, so it is the sharp interval envelope for this complete-data procedure and these per-pair constraints. This is not a claim of optimality among confidence procedures. Missingness may depend on outcomes: coverage follows from pointwise containment without a missing-at-random assumption. All-unknown data give  $[-1, 1]$ . Complete-case deletion and response-dependent retries are not part of the procedure.

**Practical decision.** With  $\varepsilon = .05$ , declare the prespecified within-margin result only when the unrounded interval lies strictly inside  $(-\varepsilon, \varepsilon)$ . Declare an effect beyond the positive or negative margin only when  $L > \varepsilon$  or  $U < -\varepsilon$ , respectively. Otherwise the precision assessment is inconclusive. Equality with a margin does not pass a strict criterion. A nonsignificant McNemar test is not a within-margin result. The exact McNemar null additionally requires fair discordant directions; it is not an exact test of arbitrary heterogeneous group means that merely average to zero.

**Extensions not justified by this argument.** The proof requires a fixed  $N$ , conditional group independence, and a prespecified common  $\lambda$  within every product. For example, with two deterministic differences  $(1, -1)$  and  $m = 0$ , group-specific values  $(\lambda_1, \lambda_2) = (.5, .1)$  give a product of 1.35 under a true average-zero null. For one fair  $D \in \{-1, 1\}$ , selecting the larger of the  $\lambda = .1$  and  $.5$  products after observing  $D$  has expectation 1.2. If two group differences are the same fair random sign, their common  $\lambda = .5$  product has expectation 1.25. Finally, for a planned  $N = 22$  deterministic sequence consisting of eleven  $+1$  values followed by eleven  $-1$  values, the planned average is zero, but the partial product after the first eleven observations is  $1.5^{11} > 80$ . These examples exclude group-specific bets, post-outcome maximization, dependent groups, and optional stopping from the stated guarantee. They do not limit other betting methods proved valid under different assumptions.

## O ORIENTATION OMISSION BOUNDS AND CONDITIONAL INFERENCE

Crossed evaluation has established predecessors, including Winoground (Thrush et al., 2022). The following argument specifies the classes and inference targets of the orientation assay; crossing and the statistical inequalities are not claimed as new principles.

**Restricted-input classes.** Fix a source group and layout, with ordered candidate-image bytes  $C$  and four distinct correct labels. The inputs  $(U, CW)$ ,  $(U, CCW)$ ,  $(RU, CW)$ ,  $(RU, CCW)$  have targets  $R^3U$ ,  $RU$ ,  $U$ ,  $R^2U$ , respectively, where  $R$  is a counterclockwise quarter turn. A rule omits a factor when its response marginal is unchanged within each pair differing only in that factor. Invalid responses form an additional output symbol and receive zero correctness.

**Proposition (orientation omission ceilings).** A rule omitting either initial orientation or rotation direction has expected quartet accuracy at most  $1/2$ . A candidate-only rule has expected accuracy at most  $1/4$ , with equality when it always returns a valid label. For deterministic rules the corresponding bounds hold for realized quartet accuracy. These statements require no independence among responses within the quartet.

**Proof.** For either omitted factor, partition the four questions into two pairs with identical retained inputs. Each pair has distinct correct labels  $y \neq y'$ , while both output marginals equal some distribution  $q$ . The expected number of correct answers in that pair is  $q(y) + q(y') \leq 1$ . Sum the two bounds and divide by four. For a candidate-only rule all four response marginals agree, so expected quartet accuracy is  $\frac{1}{4} \sum_{y \in \{A, B, C, D\}} q(y) \leq 1/4$ . Only marginal probabilities enter the proof. Deterministic rules are the point-mass special case.  $\square$

The two stored RGB initial images have identical exact foreground area and centroid, unchanged by the explicit 1344-pixel preprocessing stage. Any rule whose initial-image representation contains only these quantities remains in the initial-omission class, even if it uses the instruction and all candidate pixels. Its ceiling is  $1/2$ . The literal centroid executor is more restricted: rotating about the unchanged center produces the same point for both instructions, and its nearest-candidate prediction is candidate-only, with quartet accuracy  $1/4$ . A method that additionally reads initial shape features

need not belong to either class. A statistically significant excess rejects the stated classes under the assumptions below; it does not establish human-like mental rotation or exclude strategies based on other shape cues. Equality of learned center estimates or centroids after unknown provider-native processing is not established by these raw-image identities.

**Crossed tests conditional on the task set.** Let  $X_g \in [0, 1]$  be group  $g$ 's mean correctness over its four questions. Condition on the frozen task set and assume stable restricted-input response marginals and independent response randomness between groups. Within-group dependence is unrestricted. Under either omission null,  $\mathbb{E}X_g \leq 1/2$ , and the conservative Hoeffding p-value bound is

$$p_H = \exp\{-2N[\max(0, \bar{X}_L - 1/2)]^2\}, \quad N = 128.$$

The full-planned lower accuracy  $\bar{X}_L$  gives unknown outcomes their adverse value solely for computing this bound; they remain distinct from observed errors in reporting. Decreasing the statistic cannot make the test anti-conservative. Groups need not have identical distributions. Stateless, shuffled calls reduce, but do not prove the absence of, service drift, metadata side channels or shared response randomness.

**Reference matched test and unknown responses.** For the reference question set  $D_g = Y_{g,I} - Y_{g,S} \in \{-1, 0, 1\}$ , with positive values denoting an improvement under intervention. If all outcomes are observed, let  $w, l$  count improvements and deteriorations. The exact two-sided McNemar p value is

$$p_M(w, l) = \min\{1, 2 \Pr[\text{Bin}(w + l, 1/2) \leq \min(w, l)]\},$$

and equals one when  $w + l = 0$ . Its null assumes independent paired groups with fair discordant directions, as in the usual iid matched-pair model. It is not an exact test of arbitrary heterogeneous effects merely averaging to zero.

With unknown responses, a dynamic program tracks every feasible pair of counts  $(w, l)$  over the full planned sample, respecting all known correctness values. Both-unknown pairs allow  $D_g \in \{-1, 0, 1\}$ ; a partially known pair allows only the two differences consistent with its known member. The primary p value is the maximum  $p_M$  over these completions. It pointwise dominates the p value from the conceptual complete planned responses, and is therefore conservative under the complete-data null even if missingness depends on outcomes. Completed invalid answers are fixed errors, not unknown values. The ordinary complete-pair subset test is secondary.

Holm correction at .05 covers the twelve crossed tests and four reference tests together. It requires valid marginal p values but not independence among tests. No favorable subset or secondary contrast replaces this family.

**A separate simultaneous precision family.** Define  $W_g = \mathbf{1}\{D_g = 1\}$  and  $L_g = \mathbf{1}\{D_g = -1\}$ , with average conditional probabilities  $\bar{p}_+$  and  $\bar{p}_-$ . The fixed-task target is  $\Delta = N^{-1} \sum_g \mathbb{E}D_g = \bar{p}_+ - \bar{p}_-$ . For independent Bernoulli variables  $B_g$  with possibly different probabilities  $p_g$ , concavity of the logarithm gives, for every real  $t$ ,

$$\begin{aligned} \mathbb{E} \exp\left(t \sum_g B_g\right) &= \prod_g (1 - p_g + p_g e^t) \\ &\leq (1 - \bar{p} + \bar{p} e^t)^N. \end{aligned}$$

Chernoff optimization bounds either corresponding tail at empirical proportion  $q$  by  $\exp[-N \text{kl}(q||\bar{p})]$ , where

$$\text{kl}(q||p) = q \log(q/p) + (1 - q) \log[(1 - q)/(1 - p)].$$

Let  $\ell(q), u(q)$  invert  $N \text{kl}(q||p) = \log(1/\eta)$  on the two sides of  $q$ , with  $\eta = .05/16$ . Boundary conventions are  $\ell(0) = 0$ ,  $u(1) = 1$ ,  $u(0) = 1 - \eta^{1/N}$  and  $\ell(1) = \eta^{1/N}$ . Applying the two one-sided bounds to both indicators for each of four models and taking a union bound gives simultaneous coverage at least .95 for the four intervals

$$[\ell(w/N) - u(l/N), u(w/N) - \ell(l/N)].$$

Dependence between  $W_g$  and  $L_g$ , or between models, does not affect the union bound; independence across source groups is required. The argument does not assume an iid binomial count distribution.

For  $v$  unknown reference pairs, use the conservative completion envelope

$$[\ell(w/N) - u((l+v)/N), u((w+v)/N) - \ell(l/N)],$$

clipped to  $[-1, 1]$ . Monotonicity of the endpoints ensures that this contains the interval for every possible completion, including when partially known pairs restrict its feasible sign further. The predeclared practical tolerance is five percentage points: only strict containment of the unrounded interval in  $(-.05, .05)$  establishes that precision statement. With all differences zero and none unknown, its radius is  $1 - (.003125)^{1/128} \approx .0440647$ , rather than zero. Precision at an accuracy floor does not establish competence. This four-interval family and the 16-test discovery family have separate error guarantees, not a joint 95% guarantee across both families. Corrections also remain specific to each frozen study: no single familywise guarantee is asserted across all exploratory and prospective statements in the paper.

**Sensitivity and secondary contrasts.** At the first Holm threshold  $.05/16$ , the crossed bound requires at least 333/512 correct answers. For one-directional reference discordance, ten pairs yield exact  $p = .001953125$ , whereas nine yield  $.00390625$ ; this minimum significant configuration is not a power guarantee. In an illustrative iid trinomial scenario with discordance probability  $q$ ,  $M \sim \text{Bin}(128, q)$  and  $w \mid M \sim \text{Bin}(M, (q + \Delta)/(2q))$ . Summing the exact rejection probabilities at  $.003125$  gives 80% power at approximately 9.86, 14.66, 21.22, 30.22 percentage-point absolute effects for  $q = .1, .2, .4, .8$ , respectively. These prospective scenarios do not assert that actual provider outcomes are iid.

Reference and nonreference masks are fixed before responses. Their complete group effects satisfy  $\Delta_{\text{all}} = (\Delta_{\text{ref}} + 3\Delta_{\text{nonref}})/4$ . Their interaction and whole-quartet layout contrasts are descriptive, with 10,000 whole-group bootstrap resamples (seed 202609171803), retaining all tasks and layouts together. Incomplete groups keep full-plan identification bounds, and any complete-group subset is labeled. Neither a nonsignificant aggregate contrast nor a zero-width bootstrap establishes equivalence.

## P BITMAP CONSTRUCTION, GEOMETRIC CONTROLS AND FROZEN FILTERS

The orientation generator is an independent PIL/NumPy implementation; it does not import STARE rendering or transformation primitives. This appendix states the pre-response construction and acceptance rules, rather than additional model results.

**Equal-mass integer construction.** Draw tile count  $n$  uniformly from  $\{5, \dots, 9\}$ . Starting from one lattice cell, repeatedly select uniformly from the sorted vacant four-neighbor frontier until there are  $n$  edge-connected cells  $q_1, \dots, q_n \in \mathbb{Z}^2$ . Reject shapes invariant under any nontrivial quarter turn, allowing translation when comparing lattice sets. Use square tiles of side  $t = 4n$  pixels with centers

$$z_i = 4nq_i - 4 \sum_{j=1}^n q_j.$$

These centers are integers and sum to zero. On the 512-by-512 plot, the pixel center is  $(255.5, 255.5)$ . Tiles have identical mass  $t^2$ , touch along edges without seams and do not overlap, so the foreground has exact area  $nt^2$  and that exact centroid. Each orientation is checked against both integer coordinate rotation and exact bitmap quarter-turn execution. Tile identity is retained when rotating the centered coordinates.

The foreground is RGB(35, 111, 185), painted after the neutral grid and origin so those marks cannot erase it. Each plot has a 32-pixel label strip above it. Four candidate panels form a 1024-by-1088 montage, A/B above C/D; initial images are 512-by-544. The public full-panel crop boxes, written as  $(x_0, y_0, x_1, y_1)$ , are  $(0, 0, 512, 544)$ ,  $(512, 0, 1024, 544)$ ,  $(0, 544, 512, 1088)$  and  $(512, 544, 1024, 1088)$ . The explicit RGB maximum-edge 1344 thumbnail operation must preserve dimensions and pixels exactly. This statement does not cover unknown provider-native image processing.

**Rational matched transport.** For each proposal, independently draw the reference task index uniformly from four possibilities, candidate-label permutation, foil-to- $k$  assignment,  $s \in \{-1, 1\}$  and both coordinates of  $c$  from  $\{-50, -25, 25, 50\}$ . Let  $J$  denote a counterclockwise quarter turn and  $b = skJc$  a foil center. The sham and intervention use

$$R_{\pm} = \frac{3}{5}I \pm s\frac{4}{5}J, \quad b_{\pm} = c + R_{\pm}(b - c).$$

Expanding yields

$$\begin{aligned} b_+ &= (.4 - .8k)c + s(.6k - .8)Jc, \\ b_- &= (.4 + .8k)c + s(.6k + .8)Jc. \end{aligned}$$

For  $k = m/5$  their coefficients have denominator 25; since  $c$  has coordinates divisible by 25, every center and translation is integral. Orthogonality of  $c$  and  $Jc$  gives

$$\frac{\|b_+\|^2}{\|c\|^2} = .8 - 1.6k + k^2 < 1, \quad \frac{\|b_-\|^2}{\|c\|^2} = .8 + 1.6k + k^2 > 1$$

for each  $k \in \{1/5, 2/5, 3/5\}$ . The original ratio is  $k^2 < 1$ . Thus the reference candidate is uniquely farthest in original/sham and uniquely nearest in intervention.

All centers undergo one common rigid transformation about  $c$  in either condition, with the reference point fixed. Hence all pairwise centroid distances are preserved. Each foil moves by the same amount in the two conditions because  $\|(R_{\pm} - I)(b - c)\|^2 = 2(1 - 3/5)\|b - c\|^2$ . For indexed tile centers  $x_{ij} = c_i + u_{ij}$  with  $\sum_j u_{ij} = 0$ , expansion gives

$$\frac{1}{n} \sum_j \|x_{ij} - x_{\ell j}\|^2 = \|c_i - c_{\ell}\|^2 + \frac{1}{n} \sum_j \|u_{ij} - u_{\ell j}\|^2.$$

The first term is rotation invariant; the second is unchanged because each candidate’s centered orientation is only translated. This proves the indexed equal-mass tile-center MSE invariant even for different candidate orientations. It is not a claim about arbitrary silhouette distances or perceptual difficulty.

The same three candidate images support all four crossed questions. Only the reference question has its correct panel fixed across layouts; every other correct orientation moves. A relational position shortcut survives: normalized sums of squared centroid distances are 3.56 for the reference and 1.24, 1.24, 1.56 for the three original foils, and remain unchanged by transport. It can therefore select the reference in all layouts while scoring only one question per full quartet. A pure radial rule likewise scores one per quartet, although its reference correctness changes from one to zero between sham and intervention. These are construction properties, not empirical claims about model strategy.

**Frozen structural filters and split.** Accept only proposals passing the following pre-response checks: distinct tiles and orientations; four-connected foreground; exact area and centroid; independent coordinate/bitmap execution agreement; no clipping; at least 12 pixels raw boundary clearance; the geometric identities above; unique exact pixel-oracle answer agreeing with source execution; and unchanged reference RGB panel and explicit preprocessing. The configuration also requires colored extent at least 12 pixels, relative area error at most .05, centroid drift at most 1 pixel, no loss during centroid alignment, correct aligned-mask IoU at least .9 and a gap of at least .05 to the next-best orientation. These are structural visibility and uniqueness gates, not model-accuracy gates. Several are redundant under exact preprocessing identity; the IoU gap can nevertheless condition the distribution of shape difficulty.

A nonprivileged pixel oracle extracts foreground from the actual initial and candidate RGB images, parses the displayed direction, rotates the initial mask, aligns masks by translation and selects the unique exact match. It reads no target-orientation or answer metadata when predicting. The privileged source oracle instead has access to generated tile/program data. The perfect solvability requirement is built into acceptance; its satisfaction is not a learned-model performance improvement. Radial, relational and literal centroid-executor outcomes are recorded separately.

For calibration, collect four distinct shape rotation orbits. Define an orbit by taking the lexicographic minimum of the normalized sorted lattice tile set over all four quarter turns. Freeze those four orbits

as the formal exclusion registry; changing labels or positions does not permit reuse. Formal proposals are accepted independently subject to this fixed registry and the geometric filters. Repeated formal orbits are retained and reported, without adaptive deduplication or imposed marginal balancing. The resulting distribution is not uniform over shape classes. Acceptance may alter the marginal distribution of variables initially sampled uniformly, including the reference task index.

Formal generation targets 128 accepted groups with a 100,000-proposal cap and records every rejection, stopping without relaxing thresholds if unfilled. Generation seeds are 202609171800 for calibration and 202609171801 for formal data; shuffled request-order seeds are 202609171799 and 202609171802, respectively. Every group has 12 full-input and three candidate-only requests. All four model configurations receive every calibration request; at least 57 of 60 valid parsed outputs per configuration is the prespecified interface criterion. Scientific calibration accuracy is reported separately and does not select models, prompts, shapes or formal sample size. Floor or ceiling accuracy is not a reason to redesign an otherwise valid assay. There are no per-response retries, repairs or application-level response caches. Source, configuration, prompts, images and analysis dependencies are hash bound before formal responses; unknown or unattempted responses retain planned denominators.

## Q COMPLETE ORIENTATION-ASSAY RESULTS

The tables in this appendix are generated from the frozen analysis outputs. They retain every pre-specified model, layout and comparison. O, S and I denote original, sham and intervention. Counts of model requests, derived label evaluations and independent source groups are kept distinct. Intervals based on missing information, descriptive bootstrap intervals, bounded-group intervals and the simultaneous precision family answer different questions.

### ALL CO-PRIMARY TESTS.

| Model           | Hypothesis    | Full-plan bounds | Raw $p$                | Holm $p$               | Decision      |
|-----------------|---------------|------------------|------------------------|------------------------|---------------|
| Gemma 3 27B     | Crossed O     | [25.6, 25.6]     | 1.000                  | 1.000                  | Do not reject |
| Gemma 3 27B     | Crossed S     | [24.2, 24.2]     | 1.000                  | 1.000                  | Do not reject |
| Gemma 3 27B     | Crossed I     | [27.7, 27.7]     | 1.000                  | 1.000                  | Do not reject |
| Gemma 3 27B     | Reference I-S | [+10.2, +10.2]   | 0.029                  | 0.383                  | Do not reject |
| Mistral 3.1 24B | Crossed O     | [27.9, 27.9]     | 1.000                  | 1.000                  | Do not reject |
| Mistral 3.1 24B | Crossed S     | [29.5, 29.5]     | 1.000                  | 1.000                  | Do not reject |
| Mistral 3.1 24B | Crossed I     | [28.7, 28.7]     | 1.000                  | 1.000                  | Do not reject |
| Mistral 3.1 24B | Reference I-S | [+0.0, +0.0]     | 1.000                  | 1.000                  | Do not reject |
| Qwen2.5-VL 3B   | Crossed O     | [24.4, 24.4]     | 1.000                  | 1.000                  | Do not reject |
| Qwen2.5-VL 3B   | Crossed S     | [24.2, 24.2]     | 1.000                  | 1.000                  | Do not reject |
| Qwen2.5-VL 3B   | Crossed I     | [23.8, 23.8]     | 1.000                  | 1.000                  | Do not reject |
| Qwen2.5-VL 3B   | Reference I-S | [+5.5, +5.5]     | 0.039                  | 0.469                  | Do not reject |
| GPT-5.6 Sol     | Crossed O     | [93.2, 93.2]     | $1.93 \times 10^{-21}$ | $2.90 \times 10^{-20}$ | Reject        |
| GPT-5.6 Sol     | Crossed S     | [93.0, 93.0]     | $2.97 \times 10^{-21}$ | $4.16 \times 10^{-20}$ | Reject        |
| GPT-5.6 Sol     | Crossed I     | [93.8, 93.8]     | $5.24 \times 10^{-22}$ | $8.39 \times 10^{-21}$ | Reject        |
| GPT-5.6 Sol     | Reference I-S | [+3.1, +3.1]     | 0.219                  | 1.000                  | Do not reject |

Twelve one-sided crossed-accuracy tests and four two-sided reference contrasts form one Holm family of 16. Crossed bounds are percentages; reference bounds are percentage points. Crossed  $p$  is a bounded-group Hoeffding bound. Reference exactness requires independent matched groups with fair discordant directions; unknown outcomes use the completion-supremum  $p$ .

### REFERENCE PRECISION AND DISCORDANCE.

| Model           | Wins | Losses | Both correct | Both wrong | Unknown | Win probability % | Loss probability % | $\Delta$ CI pp |
|-----------------|------|--------|--------------|------------|---------|-------------------|--------------------|----------------|
| Gemma 3 27B     | 22   | 9      | 16           | 81         | 0       | [7.9, 30.3]       | [1.8, 17.2]        | [-9.3, +28.5]  |
| Mistral 3.1 24B | 13   | 13     | 24           | 78         | 0       | [3.4, 21.5]       | [3.4, 21.5]        | [-18.0, +18.0] |
| Qwen2.5-VL 3B   | 8    | 1      | 29           | 90         | 0       | [1.4, 16.1]       | [0.0, 6.8]         | [-5.4, +16.1]  |
| GPT-5.6 Sol     | 5    | 1      | 121          | 1          | 0       | [0.5, 12.6]       | [0.0, 6.8]         | [-6.3, +12.6]  |

All 128 planned reference pairs remain in the Bernoulli KL inversions. Sixteen one-sided tails, each  $\delta = .003125$ , cover the four effect intervals simultaneously at at least 95% under independent possibly heterogeneous groups. Unknown-pair upper counts are adverse completion envelopes. This precision family is separate from primary discovery testing.

#### CROSSED UNCERTAINTY.

| Model           | Layout | Complete groups | Full-plan bounds % | Bootstrap 95% % | One-sided lower % | Bounded 95% % |
|-----------------|--------|-----------------|--------------------|-----------------|-------------------|---------------|
| Gemma 3 27B     | O      | 128/128         | [25.6, 25.6]       | [22.9, 28.5]    | 14.8              | [13.6, 37.6]  |
| Gemma 3 27B     | S      | 128/128         | [24.2, 24.2]       | [21.7, 26.8]    | 13.4              | [12.2, 36.2]  |
| Gemma 3 27B     | I      | 128/128         | [27.7, 27.7]       | [24.6, 31.1]    | 16.9              | [15.7, 39.7]  |
| Mistral 3.1 24B | O      | 128/128         | [27.9, 27.9]       | [25.2, 30.7]    | 17.1              | [15.9, 39.9]  |
| Mistral 3.1 24B | S      | 128/128         | [29.5, 29.5]       | [27.1, 32.0]    | 18.7              | [17.5, 41.5]  |
| Mistral 3.1 24B | I      | 128/128         | [28.7, 28.7]       | [26.0, 31.4]    | 17.9              | [16.7, 40.7]  |
| Qwen2.5-VL 3B   | O      | 128/128         | [24.4, 24.4]       | [21.9, 27.0]    | 13.6              | [12.4, 36.4]  |
| Qwen2.5-VL 3B   | S      | 128/128         | [24.2, 24.2]       | [21.7, 26.8]    | 13.4              | [12.2, 36.2]  |
| Qwen2.5-VL 3B   | I      | 128/128         | [23.8, 23.8]       | [21.3, 26.4]    | 13.0              | [11.8, 35.8]  |
| GPT-5.6 Sol     | O      | 128/128         | [93.2, 93.2]       | [90.6, 95.5]    | 82.3              | [81.2, 100.0] |
| GPT-5.6 Sol     | S      | 128/128         | [93.0, 93.0]       | [90.6, 95.1]    | 82.2              | [81.0, 100.0] |
| GPT-5.6 Sol     | I      | 128/128         | [93.8, 93.8]       | [91.0, 96.1]    | 82.9              | [81.7, 100.0] |

Full-plan brackets bound unknown outcomes. Bootstrap intervals use complete source groups, 10,000 draws and seed 202609171803, and are descriptive/unadjusted. Bounded intervals target the full fixed task set under independent bounded groups. A degenerate bootstrap is not an equivalence statement.

#### QUARTET CORRECTNESS PATTERNS.

| Model           | Layout | $K = 0$ | $K = 1$ | $K = 2$ | $K = 3$ | $K = 4$ | Incomplete | All-four bounds |
|-----------------|--------|---------|---------|---------|---------|---------|------------|-----------------|
| Gemma 3 27B     | O      | 24      | 79      | 23      | 2       | 0       | 0          | 0/128           |
| Gemma 3 27B     | S      | 24      | 85      | 18      | 1       | 0       | 0          | 0/128           |
| Gemma 3 27B     | I      | 23      | 73      | 27      | 5       | 0       | 0          | 0/128           |
| Mistral 3.1 24B | O      | 16      | 84      | 25      | 3       | 0       | 0          | 0/128           |
| Mistral 3.1 24B | S      | 8       | 92      | 25      | 3       | 0       | 0          | 0/128           |
| Mistral 3.1 24B | I      | 13      | 86      | 26      | 3       | 0       | 0          | 0/128           |
| Qwen2.5-VL 3B   | O      | 24      | 83      | 21      | 0       | 0       | 0          | 0/128           |
| Qwen2.5-VL 3B   | S      | 23      | 86      | 19      | 0       | 0       | 0          | 0/128           |
| Qwen2.5-VL 3B   | I      | 23      | 89      | 15      | 1       | 0       | 0          | 0/128           |
| GPT-5.6 Sol     | O      | 0       | 1       | 5       | 22      | 100     | 0          | 100/128         |
| GPT-5.6 Sol     | S      | 0       | 0       | 5       | 26      | 97      | 0          | 97/128          |
| GPT-5.6 Sol     | I      | 0       | 2       | 4       | 18      | 104     | 0          | 104/128         |

$K$  is correct answers within a four-question quartet. Known invalid responses count as wrong.  $K$  bins include complete groups only; bins plus incomplete groups sum to 128. Individual incomplete-group  $K$  bounds are retained in the table-data JSON.

#### REFERENCE AND COMPLEMENTARY QUESTION MASKS.

The designated reference question is chosen before outcomes. Its correct panel remains fixed under the intervention; the other three correct cor-

| els             | move.  | Averaging   | these           | masks  | cannot  | establish       | reference       | equivalence. |
|-----------------|--------|-------------|-----------------|--------|---------|-----------------|-----------------|--------------|
| Model           | Layout | Fixed mask  | Correct/planned | bounds | Unknown | Complete groups | Bootstrap 95% % |              |
| Gemma 3 27B     | O      | Reference   | 25/128          |        | 0       | 128/128         | [12.5, 26.6]    |              |
| Gemma 3 27B     | O      | Other three | 106/384         |        | 0       | 128/128         | [24.0, 31.0]    |              |
| Gemma 3 27B     | S      | Reference   | 25/128          |        | 0       | 128/128         | [13.3, 26.6]    |              |
| Gemma 3 27B     | S      | Other three | 99/384          |        | 0       | 128/128         | [22.4, 29.2]    |              |
| Gemma 3 27B     | I      | Reference   | 38/128          |        | 0       | 128/128         | [21.9, 37.5]    |              |
| Gemma 3 27B     | I      | Other three | 104/384         |        | 0       | 128/128         | [23.2, 31.2]    |              |
| Mistral 3.1 24B | O      | Reference   | 32/128          |        | 0       | 128/128         | [18.0, 32.8]    |              |
| Mistral 3.1 24B | O      | Other three | 111/384         |        | 0       | 128/128         | [25.5, 32.0]    |              |
| Mistral 3.1 24B | S      | Reference   | 37/128          |        | 0       | 128/128         | [21.1, 36.7]    |              |
| Mistral 3.1 24B | S      | Other three | 114/384         |        | 0       | 128/128         | [26.3, 33.1]    |              |
| Mistral 3.1 24B | I      | Reference   | 37/128          |        | 0       | 128/128         | [21.1, 36.7]    |              |
| Mistral 3.1 24B | I      | Other three | 110/384         |        | 0       | 128/128         | [25.0, 32.6]    |              |
| Qwen2.5-VL 3B   | O      | Reference   | 29/128          |        | 0       | 128/128         | [15.6, 30.5]    |              |
| Qwen2.5-VL 3B   | O      | Other three | 96/384          |        | 0       | 128/128         | [21.6, 28.4]    |              |
| Qwen2.5-VL 3B   | S      | Reference   | 30/128          |        | 0       | 128/128         | [16.4, 31.2]    |              |
| Qwen2.5-VL 3B   | S      | Other three | 94/384          |        | 0       | 128/128         | [21.1, 27.9]    |              |
| Qwen2.5-VL 3B   | I      | Reference   | 37/128          |        | 0       | 128/128         | [21.1, 36.7]    |              |
| Qwen2.5-VL 3B   | I      | Other three | 85/384          |        | 0       | 128/128         | [18.7, 25.8]    |              |
| GPT-5.6 Sol     | O      | Reference   | 120/128         |        | 0       | 128/128         | [89.1, 97.7]    |              |
| GPT-5.6 Sol     | O      | Other three | 357/384         |        | 0       | 128/128         | [90.1, 95.6]    |              |
| GPT-5.6 Sol     | S      | Reference   | 122/128         |        | 0       | 128/128         | [91.4, 98.4]    |              |
| GPT-5.6 Sol     | S      | Other three | 354/384         |        | 0       | 128/128         | [89.3, 94.8]    |              |
| GPT-5.6 Sol     | I      | Reference   | 126/128         |        | 0       | 128/128         | [96.1, 100.0]   |              |
| GPT-5.6 Sol     | I      | Other three | 354/384         |        | 0       | 128/128         | [88.8, 95.3]    |              |

The reference question and its three-question complement are selected before outcomes. Planned denominators are 128 and 384 respectively. Bootstrap intervals resample complete source groups and are descriptive; unknown bounds retain the full population.

#### INTERVENTION MINUS SHAM BY FIXED MASK.

| Model           | Fixed mask              | Complete groups | Full effect bounds pp | Complete-group mean pp | Bootstrap 95% pp |
|-----------------|-------------------------|-----------------|-----------------------|------------------------|------------------|
| Gemma 3 27B     | All four                | 128/128         | [+3.5, +3.5]          | +3.5                   | [-0.2, +7.2]     |
| Gemma 3 27B     | Reference               | 128/128         | [+10.2, +10.2]        | +10.2                  | [+2.3, +18.8]    |
| Gemma 3 27B     | Other three             | 128/128         | [+1.3, +1.3]          | +1.3                   | [-3.4, +5.7]     |
| Gemma 3 27B     | Reference — other three | 128/128         | [+8.9, +8.9]          | +8.9                   | [-1.0, +19.0]    |
| Mistral 3.1 24B | All four                | 128/128         | [-0.8, -0.8]          | -0.8                   | [-4.1, +2.7]     |
| Mistral 3.1 24B | Reference               | 128/128         | [+0.0, +0.0]          | +0.0                   | [-7.8, +7.8]     |
| Mistral 3.1 24B | Other three             | 128/128         | [-1.0, -1.0]          | -1.0                   | [-5.2, +2.9]     |
| Mistral 3.1 24B | Reference — other three | 128/128         | [+1.0, +1.0]          | +1.0                   | [-8.1, +10.2]    |
| Qwen2.5-VL 3B   | All four                | 128/128         | [-0.4, -0.4]          | -0.4                   | [-2.7, +2.1]     |
| Qwen2.5-VL 3B   | Reference               | 128/128         | [+5.5, +5.5]          | +5.5                   | [+1.6, +10.2]    |
| Qwen2.5-VL 3B   | Other three             | 128/128         | [-2.3, -2.3]          | -2.3                   | [-5.5, +0.8]     |
| Qwen2.5-VL 3B   | Reference — other three | 128/128         | [+7.8, +7.8]          | +7.8                   | [+2.3, +13.8]    |
| GPT-5.6 Sol     | All four                | 128/128         | [+0.8, +0.8]          | +0.8                   | [-2.3, +3.7]     |
| GPT-5.6 Sol     | Reference               | 128/128         | [+3.1, +3.1]          | +3.1                   | [+0.0, +7.0]     |
| GPT-5.6 Sol     | Other three             | 128/128         | [-0.0, -0.0]          | -0.0                   | [-3.9, +3.6]     |
| GPT-5.6 Sol     | Reference — other three | 128/128         | [+3.1, +3.1]          | +3.1                   | [-2.1, +8.6]     |

intervention-sham. Whole-source-group resampling preserves paired questions. The identity  $\Delta_{all} = (\Delta_{ref} + 3\Delta_{nonref})/4$  is checked on complete groups; the interaction is  $\Delta_{ref} - \Delta_{nonref}$ . These are descriptive contrasts, with no aggregate-equivalence or response-selected-mask claim.

## SHAM MINUS ORIGINAL BY FIXED MASK.

| Model           | Fixed mask              | Complete groups | Full effect bounds pp | Complete-group mean pp | Bootstrap 95% pp |
|-----------------|-------------------------|-----------------|-----------------------|------------------------|------------------|
| Gemma 3 27B     | All four                | 128/128         | [-1.4, -1.4]          | -1.4                   | [-4.5, +2.0]     |
| Gemma 3 27B     | Reference               | 128/128         | [+0.0, +0.0]          | +0.0                   | [-7.8, +7.8]     |
| Gemma 3 27B     | Other three             | 128/128         | [-1.8, -1.8]          | -1.8                   | [-5.7, +2.3]     |
| Gemma 3 27B     | Reference — other three | 128/128         | [+1.8, +1.8]          | +1.8                   | [-7.6, +10.9]    |
| Mistral 3.1 24B | All four                | 128/128         | [+1.6, +1.6]          | +1.6                   | [-1.8, +4.9]     |
| Mistral 3.1 24B | Reference               | 128/128         | [+3.9, +3.9]          | +3.9                   | [-3.1, +10.9]    |
| Mistral 3.1 24B | Other three             | 128/128         | [+0.8, +0.8]          | +0.8                   | [-3.1, +4.7]     |
| Mistral 3.1 24B | Reference — other three | 128/128         | [+3.1, +3.1]          | +3.1                   | [-5.2, +11.7]    |
| Qwen2.5-VL 3B   | All four                | 128/128         | [-0.2, -0.2]          | -0.2                   | [-2.9, +2.5]     |
| Qwen2.5-VL 3B   | Reference               | 128/128         | [+0.8, +0.8]          | +0.8                   | [-3.9, +5.5]     |
| Qwen2.5-VL 3B   | Other three             | 128/128         | [-0.5, -0.5]          | -0.5                   | [-3.9, +3.1]     |
| Qwen2.5-VL 3B   | Reference — other three | 128/128         | [+1.3, +1.3]          | +1.3                   | [-5.2, +7.8]     |
| GPT-5.6 Sol     | All four                | 128/128         | [-0.2, -0.2]          | -0.2                   | [-3.3, +3.1]     |
| GPT-5.6 Sol     | Reference               | 128/128         | [+1.6, +1.6]          | +1.6                   | [-3.9, +7.0]     |
| GPT-5.6 Sol     | Other three             | 128/128         | [-0.8, -0.8]          | -0.8                   | [-4.4, +2.9]     |
| GPT-5.6 Sol     | Reference — other three | 128/128         | [+2.3, +2.3]          | +2.3                   | [-3.6, +8.3]     |

sham-original. Whole-source-group resampling preserves paired questions. The identity  $\Delta_{all} = (\Delta_{ref} + 3\Delta_{nonref})/4$  is checked on complete groups; the interaction is  $\Delta_{ref} - \Delta_{nonref}$ . These are descriptive contrasts, with no aggregate-equivalence or response-selected-mask claim.

## INTERVENTION MINUS ORIGINAL BY FIXED MASK.

| Model           | Fixed mask              | Complete groups | Full effect bounds pp | Complete-group mean pp | Bootstrap 95% pp |
|-----------------|-------------------------|-----------------|-----------------------|------------------------|------------------|
| Gemma 3 27B     | All four                | 128/128         | [+2.1, +2.1]          | +2.1                   | [-1.4, +5.7]     |
| Gemma 3 27B     | Reference               | 128/128         | [+10.2, +10.2]        | +10.2                  | [+2.3, +18.0]    |
| Gemma 3 27B     | Other three             | 128/128         | [-0.5, -0.5]          | -0.5                   | [-4.9, +3.9]     |
| Gemma 3 27B     | Reference — other three | 128/128         | [+10.7, +10.7]        | +10.7                  | [+0.8, +20.6]    |
| Mistral 3.1 24B | All four                | 128/128         | [+0.8, +0.8]          | +0.8                   | [-2.9, +4.5]     |
| Mistral 3.1 24B | Reference               | 128/128         | [+3.9, +3.9]          | +3.9                   | [-4.7, +12.5]    |
| Mistral 3.1 24B | Other three             | 128/128         | [-0.3, -0.3]          | -0.3                   | [-4.2, +3.9]     |
| Mistral 3.1 24B | Reference — other three | 128/128         | [+4.2, +4.2]          | +4.2                   | [-5.5, +13.5]    |
| Qwen2.5-VL 3B   | All four                | 128/128         | [-0.6, -0.6]          | -0.6                   | [-3.5, +2.5]     |
| Qwen2.5-VL 3B   | Reference               | 128/128         | [+6.2, +6.2]          | +6.2                   | [+0.8, +11.7]    |
| Qwen2.5-VL 3B   | Other three             | 128/128         | [-2.9, -2.9]          | -2.9                   | [-6.8, +1.0]     |
| Qwen2.5-VL 3B   | Reference — other three | 128/128         | [+9.1, +9.1]          | +9.1                   | [+1.8, +16.4]    |
| GPT-5.6 Sol     | All four                | 128/128         | [+0.6, +0.6]          | +0.6                   | [-2.0, +3.1]     |
| GPT-5.6 Sol     | Reference               | 128/128         | [+4.7, +4.7]          | +4.7                   | [+0.8, +8.6]     |
| GPT-5.6 Sol     | Other three             | 128/128         | [-0.8, -0.8]          | -0.8                   | [-3.9, +2.3]     |
| GPT-5.6 Sol     | Reference — other three | 128/128         | [+5.5, +5.5]          | +5.5                   | [+0.5, +10.7]    |

intervention-original. Whole-source-group resampling preserves paired questions. The identity  $\Delta_{all} = (\Delta_{ref} + 3\Delta_{nonref})/4$  is checked on complete groups; the interaction is  $\Delta_{ref} - \Delta_{nonref}$ . These are descriptive contrasts, with no aggregate-equivalence or response-selected-mask claim.

## CANDIDATE-ONLY RESPONSES.

A valid blind response necessarily matches one of a quartet’s four labels. The resulting 25% is an analytic consequence, not evidence of spatial skill or an observed uniquely correct candidate-only answer. Reference matching describes a preference for the designated reference candidate.

| Model           | Layout | Valid | Invalid | Unknown | Derived quartet % | Reference preference |
|-----------------|--------|-------|---------|---------|-------------------|----------------------|
| Gemma 3 27B     | O      | 128   | 0       | 0       | [25.0, 25.0]      | 34/128               |
| Gemma 3 27B     | S      | 128   | 0       | 0       | [25.0, 25.0]      | 34/128               |
| Gemma 3 27B     | I      | 128   | 0       | 0       | [25.0, 25.0]      | 33/128               |
| Mistral 3.1 24B | O      | 128   | 0       | 0       | [25.0, 25.0]      | 39/128               |
| Mistral 3.1 24B | S      | 128   | 0       | 0       | [25.0, 25.0]      | 39/128               |
| Mistral 3.1 24B | I      | 128   | 0       | 0       | [25.0, 25.0]      | 26/128               |
| Qwen2.5-VL 3B   | O      | 128   | 0       | 0       | [25.0, 25.0]      | 20/128               |
| Qwen2.5-VL 3B   | S      | 128   | 0       | 0       | [25.0, 25.0]      | 22/128               |
| Qwen2.5-VL 3B   | I      | 128   | 0       | 0       | [25.0, 25.0]      | 33/128               |
| GPT-5.6 Sol     | O      | 128   | 0       | 0       | [25.0, 25.0]      | 54/128               |
| GPT-5.6 Sol     | S      | 128   | 0       | 0       | [25.0, 25.0]      | 47/128               |
| GPT-5.6 Sol     | I      | 128   | 0       | 0       | [25.0, 25.0]      | 67/128               |

Exactly 128 candidate-only requests per model/layout, each scored analytically against its four labels. A valid answer gives 25%, invalid 0%, and unknown [0, 25]%. Reference matching is a separate preference diagnostic, not correctness for a uniquely answerable candidate-only question. No extra requests or independent samples are created.

#### CHANGES OF REQUIRED FACTORS.

| Model           | Layout | Changed factor | Switch/valid | Both correct | Unknown pairs | Invalid pairs | Switch CI %   |
|-----------------|--------|----------------|--------------|--------------|---------------|---------------|---------------|
| Gemma 3 27B     | O      | Initial state  | 77/256       | 8/256        | 0 0           |               | [24.6, 35.5]  |
| Gemma 3 27B     | O      | Direction      | 113/256      | 13/256       | 0 0           |               | [37.5, 50.8]  |
| Gemma 3 27B     | S      | Initial state  | 88/256       | 5/256        | 0 0           |               | [28.1, 40.6]  |
| Gemma 3 27B     | S      | Direction      | 119/256      | 10/256       | 0 0           |               | [39.8, 53.1]  |
| Gemma 3 27B     | I      | Initial state  | 100/256      | 9/256        | 0 0           |               | [32.8, 45.3]  |
| Gemma 3 27B     | I      | Direction      | 120/256      | 20/256       | 0 0           |               | [40.2, 53.5]  |
| Mistral 3.1 24B | O      | Initial state  | 82/256       | 12/256       | 0 0           |               | [25.8, 38.3]  |
| Mistral 3.1 24B | O      | Direction      | 81/256       | 11/256       | 0 0           |               | [25.4, 37.9]  |
| Mistral 3.1 24B | S      | Initial state  | 75/256       | 8/256        | 0 0           |               | [23.0, 35.9]  |
| Mistral 3.1 24B | S      | Direction      | 79/256       | 10/256       | 0 0           |               | [24.2, 37.5]  |
| Mistral 3.1 24B | I      | Initial state  | 70/256       | 9/256        | 0 0           |               | [21.9, 33.2]  |
| Mistral 3.1 24B | I      | Direction      | 84/256       | 10/256       | 0 0           |               | [26.6, 39.1]  |
| Qwen2.5-VL 3B   | O      | Initial state  | 53/256       | 4/256        | 0 0           |               | [16.0, 25.4]  |
| Qwen2.5-VL 3B   | O      | Direction      | 111/256      | 6/256        | 0 0           |               | [36.3, 50.8]  |
| Qwen2.5-VL 3B   | S      | Initial state  | 42/256       | 4/256        | 0 0           |               | [12.1, 21.1]  |
| Qwen2.5-VL 3B   | S      | Direction      | 108/256      | 4/256        | 0 0           |               | [34.4, 49.6]  |
| Qwen2.5-VL 3B   | I      | Initial state  | 51/256       | 4/256        | 0 0           |               | [15.2, 24.6]  |
| Qwen2.5-VL 3B   | I      | Direction      | 110/256      | 5/256        | 0 0           |               | [35.5, 50.0]  |
| GPT-5.6 Sol     | O      | Initial state  | 254/256      | 225/256      | 0 0           |               | [98.0, 100.0] |
| GPT-5.6 Sol     | O      | Direction      | 229/256      | 223/256      | 0 0           |               | [85.2, 93.4]  |
| GPT-5.6 Sol     | S      | Initial state  | 254/256      | 221/256      | 0 0           |               | [98.0, 100.0] |
| GPT-5.6 Sol     | S      | Direction      | 226/256      | 222/256      | 0 0           |               | [84.0, 92.2]  |
| GPT-5.6 Sol     | I      | Initial state  | 254/256      | 226/256      | 0 0           |               | [97.7, 100.0] |
| GPT-5.6 Sol     | I      | Direction      | 233/256      | 228/256      | 0 0           |               | [87.1, 94.5]  |

Each factor has 256 paired comparisons but only 128 independent source groups. Switch rates require two valid labels; invalid means a known pair containing an invalid answer, unknown means at least one unknown outcome. CIs resample whole groups. A changed answer alone does not establish correct composition.

#### REFERENCE RESPONSES ACROSS LAYOUTS.

| Model           | Reference layout change | Switch/valid | Known pairs | Unknown pairs | Invalid pairs |
|-----------------|-------------------------|--------------|-------------|---------------|---------------|
| Gemma 3 27B     | S to I                  | 61/128       | 128         | 0             | 0             |
| Gemma 3 27B     | O to S                  | 52/128       | 128         | 0             | 0             |
| Gemma 3 27B     | O to I                  | 58/128       | 128         | 0             | 0             |
| Mistral 3.1 24B | S to I                  | 51/128       | 128         | 0             | 0             |
| Mistral 3.1 24B | O to S                  | 49/128       | 128         | 0             | 0             |
| Mistral 3.1 24B | O to I                  | 55/128       | 128         | 0             | 0             |
| Qwen2.5-VL 3B   | S to I                  | 32/128       | 128         | 0             | 0             |
| Qwen2.5-VL 3B   | O to S                  | 32/128       | 128         | 0             | 0             |
| Qwen2.5-VL 3B   | O to I                  | 36/128       | 128         | 0             | 0             |
| GPT-5.6 Sol     | S to I                  | 6/128        | 128         | 0             | 0             |
| GPT-5.6 Sol     | O to S                  | 12/128       | 128         | 0             | 0             |
| GPT-5.6 Sol     | O to I                  | 8/128        | 128         | 0             | 0             |

All 128 reference groups remain in the transition tables. Full answer and correctness transition counts, including missing/unknown/invalid states, are preserved in table-data JSON. These response-switch summaries are descriptive.

#### PRESPECIFIED SUFFICIENT AND RESTRICTED-INPUT SOLVERS.

The source oracle uses generation metadata; the pixel-mask oracle does not. The latter’s perfect solvability is checked in construction and is not a learned-model improvement. The centroid ex-

ecutor and centroid-based omission class are different restrictions, as described in the main text.

| Rule              | Full O  | Full S  | Full I  | Blind O % / ref           | Blind S % / ref           | Blind I % / ref           |
|-------------------|---------|---------|---------|---------------------------|---------------------------|---------------------------|
| Radial (raw)      | 128/512 | 128/512 | 128/512 | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 0/128 ref   |
| Outlier (raw)     | 128/512 | 128/512 | 128/512 | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref |
| Radial (1344)     | 128/512 | 128/512 | 128/512 | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 0/128 ref   |
| Outlier (1344)    | 128/512 | 128/512 | 128/512 | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref | [25.0, 25.0]; 128/128 ref |
| Centroid executor | 128/512 | 128/512 | 128/512 | Unavailable               | Unavailable               | Unavailable               |
| Source oracle     | 512/512 | 512/512 | 512/512 | Unavailable               | Unavailable               | Unavailable               |
| Pixel-mask oracle | 512/512 | 512/512 | 512/512 | Unavailable               | Unavailable               | Unavailable               |

Full-input counts use 512 questions/layout; blind columns use one response/group and a derived quartet score. Reference preference is distinct. Comparators are all retained, with explicit availability; source oracle is privileged, pixel-mask oracle reads pixels/instruction. Predictor code is not rerun by the manuscript renderer.

#### MODEL–RULE AGREEMENT: GEMMA.

| Rule              | Full O               | Full S               | Full I               | Blind O             | Blind S             | Blind I             |
|-------------------|----------------------|----------------------|----------------------|---------------------|---------------------|---------------------|
| Radial (raw)      | 90/512 [17.6, 17.6]  | 110/512 [21.5, 21.5] | 118/512 [23.0, 23.0] | 34/128 [26.6, 26.6] | 34/128 [26.6, 26.6] | 31/128 [24.2, 24.2] |
| Outlier (raw)     | 90/512 [17.6, 17.6]  | 110/512 [21.5, 21.5] | 128/512 [25.0, 25.0] | 34/128 [26.6, 26.6] | 34/128 [26.6, 26.6] | 33/128 [25.8, 25.8] |
| Radial (1344)     | 90/512 [17.6, 17.6]  | 110/512 [21.5, 21.5] | 118/512 [23.0, 23.0] | 34/128 [26.6, 26.6] | 34/128 [26.6, 26.6] | 31/128 [24.2, 24.2] |
| Outlier (1344)    | 90/512 [17.6, 17.6]  | 110/512 [21.5, 21.5] | 128/512 [25.0, 25.0] | 34/128 [26.6, 26.6] | 34/128 [26.6, 26.6] | 33/128 [25.8, 25.8] |
| Centroid executor | 134/512 [26.2, 26.2] | 128/512 [25.0, 25.0] | 128/512 [25.0, 25.0] | Unavailable         | Unavailable         | Unavailable         |
| Source oracle     | 131/512 [25.6, 25.6] | 124/512 [24.2, 24.2] | 142/512 [27.7, 27.7] | Unavailable         | Unavailable         | Unavailable         |
| Pixel-mask oracle | 131/512 [25.6, 25.6] | 124/512 [24.2, 24.2] | 142/512 [27.7, 27.7] | Unavailable         | Unavailable         | Unavailable         |

Gemma 3 27B: agree/known outcomes with an available rule, followed by full-planned agreement bounds in percent. Known invalid responses are nonagreement; unknown outcomes remain bounded. Descriptive agreement does not identify a model mechanism.

#### MODEL–RULE AGREEMENT: MISTRAL.

| Rule              | Full O               | Full S               | Full I               | Blind O             | Blind S             | Blind I             |
|-------------------|----------------------|----------------------|----------------------|---------------------|---------------------|---------------------|
| Radial (raw)      | 114/512 [22.3, 22.3] | 120/512 [23.4, 23.4] | 147/512 [28.7, 28.7] | 39/128 [30.5, 30.5] | 39/128 [30.5, 30.5] | 31/128 [24.2, 24.2] |
| Outlier (raw)     | 114/512 [22.3, 22.3] | 120/512 [23.4, 23.4] | 118/512 [23.0, 23.0] | 39/128 [30.5, 30.5] | 39/128 [30.5, 30.5] | 26/128 [20.3, 20.3] |
| Radial (1344)     | 114/512 [22.3, 22.3] | 120/512 [23.4, 23.4] | 147/512 [28.7, 28.7] | 39/128 [30.5, 30.5] | 39/128 [30.5, 30.5] | 31/128 [24.2, 24.2] |
| Outlier (1344)    | 114/512 [22.3, 22.3] | 120/512 [23.4, 23.4] | 118/512 [23.0, 23.0] | 39/128 [30.5, 30.5] | 39/128 [30.5, 30.5] | 26/128 [20.3, 20.3] |
| Centroid executor | 117/512 [22.9, 22.9] | 117/512 [22.9, 22.9] | 118/512 [23.0, 23.0] | Unavailable         | Unavailable         | Unavailable         |
| Source oracle     | 143/512 [27.9, 27.9] | 151/512 [29.5, 29.5] | 147/512 [28.7, 28.7] | Unavailable         | Unavailable         | Unavailable         |
| Pixel-mask oracle | 143/512 [27.9, 27.9] | 151/512 [29.5, 29.5] | 147/512 [28.7, 28.7] | Unavailable         | Unavailable         | Unavailable         |

Mistral 3.1 24B: agree/known outcomes with an available rule, followed by full-planned agreement bounds in percent. Known invalid responses are nonagreement; unknown outcomes remain bounded. Descriptive agreement does not identify a model mechanism.

#### MODEL–RULE AGREEMENT: QWEN.

| Rule              | Full O               | Full S               | Full I               | Blind O             | Blind S             | Blind I             |
|-------------------|----------------------|----------------------|----------------------|---------------------|---------------------|---------------------|
| Radial (raw)      | 117/512 [22.9, 22.9] | 134/512 [26.2, 26.2] | 111/512 [21.7, 21.7] | 20/128 [15.6, 15.6] | 22/128 [17.2, 17.2] | 31/128 [24.2, 24.2] |
| Outlier (raw)     | 117/512 [22.9, 22.9] | 134/512 [26.2, 26.2] | 144/512 [28.1, 28.1] | 20/128 [15.6, 15.6] | 22/128 [17.2, 17.2] | 33/128 [25.8, 25.8] |
| Radial (1344)     | 117/512 [22.9, 22.9] | 134/512 [26.2, 26.2] | 111/512 [21.7, 21.7] | 20/128 [15.6, 15.6] | 22/128 [17.2, 17.2] | 31/128 [24.2, 24.2] |
| Outlier (1344)    | 117/512 [22.9, 22.9] | 134/512 [26.2, 26.2] | 144/512 [28.1, 28.1] | 20/128 [15.6, 15.6] | 22/128 [17.2, 17.2] | 33/128 [25.8, 25.8] |
| Centroid executor | 155/512 [30.3, 30.3] | 110/512 [21.5, 21.5] | 144/512 [28.1, 28.1] | Unavailable         | Unavailable         | Unavailable         |
| Source oracle     | 125/512 [24.4, 24.4] | 124/512 [24.2, 24.2] | 122/512 [23.8, 23.8] | Unavailable         | Unavailable         | Unavailable         |
| Pixel-mask oracle | 125/512 [24.4, 24.4] | 124/512 [24.2, 24.2] | 122/512 [23.8, 23.8] | Unavailable         | Unavailable         | Unavailable         |

Qwen2.5-VL 3B: agree/known outcomes with an available rule, followed by full-planned agreement bounds in percent. Known invalid responses are nonagreement; unknown outcomes remain bounded. Descriptive agreement does not identify a model mechanism.

#### MODEL–RULE AGREEMENT: SOL.

Agreement with a sufficient rule does not reveal a model’s internal strategy.

| Rule              | Full O               | Full S               | Full I               | Blind O             | Blind S             | Blind I             |
|-------------------|----------------------|----------------------|----------------------|---------------------|---------------------|---------------------|
| Radial (raw)      | 131/512 [25.6, 25.6] | 126/512 [24.6, 24.6] | 124/512 [24.2, 24.2] | 54/128 [42.2, 42.2] | 47/128 [36.7, 36.7] | 23/128 [18.0, 18.0] |
| Outlier (raw)     | 131/512 [25.6, 25.6] | 126/512 [24.6, 24.6] | 136/512 [26.6, 26.6] | 54/128 [42.2, 42.2] | 47/128 [36.7, 36.7] | 67/128 [52.3, 52.3] |
| Radial (1344)     | 131/512 [25.6, 25.6] | 126/512 [24.6, 24.6] | 124/512 [24.2, 24.2] | 54/128 [42.2, 42.2] | 47/128 [36.7, 36.7] | 23/128 [18.0, 18.0] |
| Outlier (1344)    | 131/512 [25.6, 25.6] | 126/512 [24.6, 24.6] | 136/512 [26.6, 26.6] | 54/128 [42.2, 42.2] | 47/128 [36.7, 36.7] | 67/128 [52.3, 52.3] |
| Centroid executor | 127/512 [24.8, 24.8] | 134/512 [26.2, 26.2] | 136/512 [26.6, 26.6] | Unavailable         | Unavailable         | Unavailable         |
| Source oracle     | 477/512 [93.2, 93.2] | 476/512 [93.0, 93.0] | 480/512 [93.8, 93.8] | Unavailable         | Unavailable         | Unavailable         |
| Pixel-mask oracle | 477/512 [93.2, 93.2] | 476/512 [93.0, 93.0] | 480/512 [93.8, 93.8] | Unavailable         | Unavailable         | Unavailable         |

GPT-5.6 Sol: agree/known outcomes with an available rule, followed by full-planned agreement bounds in percent. Known invalid responses are nonagreement; unknown outcomes remain bounded. Descriptive agreement does not identify a model mechanism.

## OTHER REFERENCE CONTRASTS.

These contrasts are descriptive and do not receive the primary family’s confirmatory interpretation.

| Model           | Reference contrast    | Known pairs | Full effect bounds pp | Complete-pair mean pp | Bootstrap 95% pp | Raw completion $p$ |
|-----------------|-----------------------|-------------|-----------------------|-----------------------|------------------|--------------------|
| Gemma 3 27B     | sham-original         | 128/128     | [+0.0, +0.0]          | +0.0                  | [-7.8, +7.8]     | 1.000              |
| Gemma 3 27B     | intervention-original | 128/128     | [+10.2, +10.2]        | +10.2                 | [+2.3, +18.0]    | 0.024              |
| Mistral 3.1 24B | sham-original         | 128/128     | [+3.9, +3.9]          | +3.9                  | [-3.1, +10.9]    | 0.405              |
| Mistral 3.1 24B | intervention-original | 128/128     | [+3.9, +3.9]          | +3.9                  | [-4.7, +12.5]    | 0.473              |
| Qwen2.5-VL 3B   | sham-original         | 128/128     | [+0.8, +0.8]          | +0.8                  | [-3.9, +5.5]     | 1.000              |
| Qwen2.5-VL 3B   | intervention-original | 128/128     | [+6.2, +6.2]          | +6.2                  | [+0.8, +11.7]    | 0.057              |
| GPT-5.6 Sol     | sham-original         | 128/128     | [+1.6, +1.6]          | +1.6                  | [-3.9, +7.0]     | 0.774              |
| GPT-5.6 Sol     | intervention-original | 128/128     | [+4.7, +4.7]          | +4.7                  | [+0.8, +8.6]     | 0.070              |

Descriptive secondary reference contrasts; no confirmatory multiplicity or precision-family claim. Bootstrap intervals use complete pairs, while bounds and completion-supremum  $p$  retain the full planned population. Non-significance is not equivalence.

## ANSWER-PATTERN ACCOUNTING.

Every four-answer pattern, including invalid and unknown states, is retained in the accompanying machine-readable CSV and analysis JSON.

| Model           | Layout | Distinct patterns | Largest pattern count | All labels valid | Invalid/unknown groups |
|-----------------|--------|-------------------|-----------------------|------------------|------------------------|
| Gemma 3 27B     | O      | 40                | 20                    | 128              | 0                      |
| Gemma 3 27B     | S      | 48                | 14                    | 128              | 0                      |
| Gemma 3 27B     | I      | 44                | 19                    | 128              | 0                      |
| Mistral 3.1 24B | O      | 39                | 50                    | 128              | 0                      |
| Mistral 3.1 24B | S      | 37                | 51                    | 128              | 0                      |
| Mistral 3.1 24B | I      | 33                | 49                    | 128              | 0                      |
| Qwen2.5-VL 3B   | O      | 27                | 52                    | 128              | 0                      |
| Qwen2.5-VL 3B   | S      | 19                | 59                    | 128              | 0                      |
| Qwen2.5-VL 3B   | I      | 23                | 50                    | 128              | 0                      |
| GPT-5.6 Sol     | O      | 50                | 9                     | 128              | 0                      |
| GPT-5.6 Sol     | S      | 49                | 7                     | 128              | 0                      |
| GPT-5.6 Sol     | I      | 46                | 8                     | 128              | 0                      |

Each pattern contains four answers in target-index order, including explicit invalid/unknown/missing states. Counts sum to 128 groups per model/layout. Every pattern frequency is retained in orientation-patterns.csv and table-data JSON; no top-pattern selection is used for inference.

## REQUEST ACCOUNTING.

| Model           | Layout | Input | Planned | Recorded | Known | Valid | Invalid | Unknown | Missing |
|-----------------|--------|-------|---------|----------|-------|-------|---------|---------|---------|
| Gemma 3 27B     | O      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Gemma 3 27B     | O      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Gemma 3 27B     | S      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Gemma 3 27B     | S      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Gemma 3 27B     | I      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Gemma 3 27B     | I      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Mistral 3.1 24B | O      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Mistral 3.1 24B | O      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Mistral 3.1 24B | S      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Mistral 3.1 24B | S      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Mistral 3.1 24B | I      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Mistral 3.1 24B | I      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | O      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | O      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | S      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | S      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | I      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| Qwen2.5-VL 3B   | I      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| GPT-5.6 Sol     | O      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| GPT-5.6 Sol     | O      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| GPT-5.6 Sol     | S      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| GPT-5.6 Sol     | S      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |
| GPT-5.6 Sol     | I      | Full  | 512     | 512      | 512   | 512   | 0       | 0       | 0       |
| GPT-5.6 Sol     | I      | Blind | 128     | 128      | 128   | 128   | 0       | 0       | 0       |

Full and blind rows count actual requests once. Unknown includes recorded transport failures and missing requests; missing is a subset of unknown. Known=valid+invalid, and planned=known+unknown=recorded+missing. Calibration is excluded.

## R CUE-OMISSION BOUNDS FOR CROSSED TASKS

Crossed image–language evaluation has established predecessors, including Winoground (Thrush et al., 2022). The following counting argument states the precise restricted-input classes tested by our two spatial constructions; we do not claim crossing as a new evaluation principle.

**Construction and omission classes.** Each group contains four tasks with the same ordered candidate-image bytes  $C$  and distinct correct labels  $y_1, \dots, y_4$ . In family A, a centered polygon  $U$  gives initial states  $S_a = U + ab$  and instructions  $Q_t = T_{tv}$ , for  $a, t \in \{-1, +1\}$ ; the candidates are the four distinct sets  $U + ab + tv$ . Holding  $t$  fixes the prompt and  $C$ ; holding  $a$  fixes the initial-image bytes and  $C$ . In family B, the initial state  $U + b$  is fixed, with  $U = -U$  as polygon sets. The four programs are  $T_v F_x, F_x T_v, T_v F_y, F_y T_v$ , composed right to left:  $F_x(x, y) = (x, -y)$ ,  $F_y(x, y) = (-x, y)$ , and  $T_v(z) = z + v$ . Writing  $p = F_x b$ , their centers are  $p + v, p + F_x v, -p + v, -p + F_y v$ , respectively. Central symmetry makes their centered polygons equal as sets. Execution with the declared rounding rule, set uniqueness, and retained-input byte identities are independently checked.

A response rule *omits a factor* if its output distribution is unchanged when that factor alone changes within a group, with the retained factors fixed. For A these pairs change  $a$  or  $t$ ; for B they change axis at fixed order or order at fixed axis. This is an invariance of the complete response distribution, not merely an absence of particular words from an explanation. The output alphabet includes an invalid/abstention symbol, always scored zero.

**Proposition (finite-group omission ceilings).** For either family, a rule omitting either of its two crossed factors has expected group accuracy at most  $1/2$ . A candidate-only rule has expected accuracy at most  $1/4$ , with equality if it always returns a valid label. For deterministic rules these bounds hold for the realized group accuracy. The expectation statements allow arbitrary dependence among the four responses within a group.

**Proof.** Partition the four tasks into the two pairs that differ only in the omitted factor. In either pair, the two correct labels  $y \neq y'$  are distinct, whereas the response marginals coincide with a distribution  $q$ . Thus the expected number of correct answers in that pair is  $q(y) + q(y') \leq 1$ . Summing the two pairs and dividing by four gives  $1/2$ . For a candidate-only rule all four marginals coincide, giving  $\frac{1}{4} \sum_{j=1}^4 q(y_j) \leq 1/4$ ; equality holds when all probability mass is on the four valid labels. Deterministic rules are point-mass special cases. Only marginal probabilities enter these calculations, so no within-group independence is needed.  $\square$

**Conditional inference and its limits.** Condition on the frozen tasks. Let  $X_g \in [0, 1]$  be the mean correctness of group  $g$ , and assume stable response marginals on its paired retained inputs and independent response randomness *between* the  $N$  groups. Under either omission class,  $\mathbb{E}[X_g] \leq 1/2$ , and Hoeffding gives the conservative one-sided p-value bound

$$p_H = \exp\left[-2N(\max\{0, \bar{X}_L - 1/2\})^2\right], \quad (24)$$

where  $\bar{X}_L$  uses the full planned denominator and assigns unknown correctness its adverse value zero solely to form a lower bound. Unknowns remain separately reported. Groups need not be identically distributed; fixed shape strata and deduplication do not justify iid generator sampling. Stateless shuffled calls reduce, but do not prove the absence of, service drift or shared response randomness. The bound assumes the stated cross-group independence and stable kernels. Holm correction permits dependence among valid p-values, including tests sharing B source groups.

A statistically significant excess over  $1/2$  rejects the corresponding omission classes under these assumptions; it does not identify an internal execution algorithm. Family B does not establish initial-image necessity. A nonsignificant result establishes neither omission nor equivalence. The separate B-matched candidate sets are outside this proposition and retain their known relational shortcut.

## S AFFINE CROSSED CONSTRUCTION IN DETAIL

The matched operator preserves the centroid-distance matrix and indexed-vertex MSEs. It therefore cannot remove a predictor based only on that invariant matrix. Moreover, a centered initial state followed by a final translation lets a solver select the corresponding final centroid without processing the initial image or previous operations. We construct two complementary four-task groups in which omitting a specified factor forces contradictory answers on identical remaining inputs. This is an application of paired compositional evaluation, as in Winoground (Thrush et al., 2022), to executable planar transformations; we do not claim crossing or joint-correctness scoring as new principles.

**Initial state and instruction (A).** For a centered polygon  $U$  and nonzero vectors  $b, v$ , let

$$S_a = U + ab, \quad Q_t = T(tv), \quad C = \{U + ab + tv : a, t \in \{-1, +1\}\}. \quad (25)$$

A single random label permutation and the same candidate-image bytes serve all four tasks. Each task displays its own initial image  $S_a$  and instruction  $Q_t$ , with no numerical description of  $b$ . We reject coincident targets. Holding  $t$  fixed, changing  $a$  changes the answer while preserving  $(Q, C)$ ; holding  $a$  fixed, changing  $t$  changes the answer while preserving  $(S, C)$ . Any deterministic solver omitting either initial image or instruction scores at most  $2/4$ ; a candidate-only solver scores at most  $1/4$ , with equality for a valid deterministic label. A randomized solver has the corresponding expectation bounds under a stable response rule.

**Axis and noncommuting order (B).** Start from an off-center polygon  $X = U + b$ , use a translation  $T_v$  with both coordinates nonzero, and let  $F_x, F_y$  reflect across the horizontal and vertical axes, respectively. The four candidate outputs are

$$C = \{T_v F_x X, F_x T_v X, T_v F_y X, F_y T_v X\}. \quad (26)$$

Composition acts right to left. All four instructions share the same initial and candidate images; only axis and order change. We require four distinct outputs and render them under one fixed label permutation. For each axis, the two orders have different answers, and for each order the axes have different answers. Omitting either factor therefore gives the same  $1/2$  ceiling, while a question-blind rule has expectation at most  $1/4$  (equality for valid labels). This family does not test initial-image

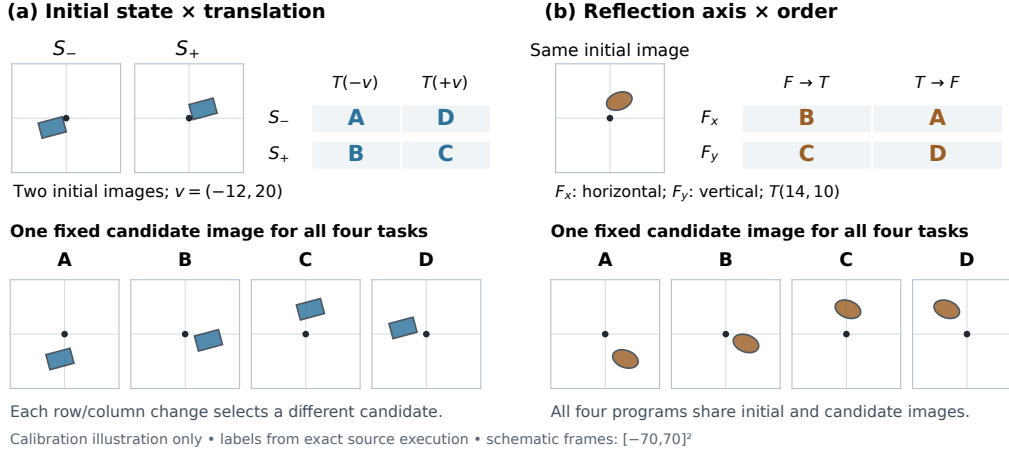


Figure 11: Fixed calibration groups illustrate the crossed design. Each row or column change makes a different candidate correct while preserving the other factor. All four tasks within a family share identical candidate-image bytes. The schematic redraw uses a tighter common frame; task rendering is unchanged.

necessity; A does. Appendix R gives the exact omission classes and proofs. A statistically significant excess over their expectation bounds rejects those classes, but does not reveal an internal reasoning trace or rule out every heuristic that uses both factors.

**Matched replication on the same B groups.** We preselect one of the four programs per B group and keep its initial image, instruction and exact correct output. A *separate* candidate set replaces the wrong outputs with copies of the correct centered shape at the centroids specified in Section 3. Original, sham and intervention then preserve the same geometric quantities as before. The balanced-target proof applies to B’s crossed set, not this separate matched set; the latter still admits the perfect relational-outlier rule. This tests the matched control on a noncommuting task construction while keeping its diagnostic boundary explicit. Comparing blocks is descriptive: their candidate sets and numbers of instructions differ.

**Frozen generation and inference.** Each family has 32 new groups, eight each of rectangles, squares, ellipses and hexagons. Their four tasks give 256 crossed questions; B’s 32 triplets add 96, for 352 requests per model. Four disjoint calibration groups supply 22 requests per model and do not enter these results. Shape parameters, seeds, finite proposal limits, labels and program allocations are fixed before model outcomes. Source execution and an independent executor agree, targets are unique, and both original and resized images pass visibility and byte-identity checks. Preparation orientation is part of the initial shape, not an unmentioned operation in the instruction. Reflections/translations do not invoke the public executor’s separate tracked-center defect.

The twelve co-primary tests comprise eight tests of crossed mean group accuracy above  $1/2$  and four exact matched intervention–sham McNemar tests. Holm correction spans all twelve. For crossed group accuracies  $Z_g \in [0, 1]$ , we use

$$p_H = \exp\{-2N[\max(0, \bar{Z}_{\text{lower}} - 1/2)]^2\}, \quad N = 32, \quad (27)$$

where unknown outcomes receive their adverse full-denominator bound, without becoming observed failures. This conservative bound conditions on the fixed task set and assumes independent response randomness across groups and a stable response procedure within the paired inputs; within-group dependence is unrestricted. It does not assume independent individual questions or iid geometric sampling. At the first Holm threshold, at least 102/128 correct answers are needed; the design targets large departures, not near-chance equivalence. McNemar’s exactness requires independent pairs and fair discordant directions under its pairwise null (or the usual iid-pair model), rather than an arbitrary heterogeneous average-zero null. Complete-group/pair bootstrap intervals (10,000 draws) are secondary, with separate full-planned unknown bounds and nondegenerate bounded-group intervals at ceilings. All settings and statistical assumptions were frozen before the formal calls.

| Model           | A: correct | A: all four | A: Holm $p$           | B: correct | B: all four | B: Holm $p$           |
|-----------------|------------|-------------|-----------------------|------------|-------------|-----------------------|
| Gemma 3 27B     | 35/128     | 0/32        | 1.000                 | 30/128     | 0/32        | 1.000                 |
| Mistral 3.1 24B | 30/128     | 0/32        | 1.000                 | 33/128     | 0/32        | 1.000                 |
| Qwen2.5-VL 3B   | 40/128     | 0/32        | 1.000                 | 37/128     | 0/32        | 1.000                 |
| GPT-5.6 Sol     | 123/128    | 27/32       | $1.49 \times 10^{-5}$ | 121/128    | 25/32       | $3.39 \times 10^{-5}$ |

Table 32: Crossed tasks: 128 questions in 32 groups per family. A tests initial state and translation; B tests reflection axis and order. “All four” requires every answer in a group to be correct. The null is expected group accuracy  $\leq 1/2$ , with Holm correction across all 12 new co-primary tests.

| Model           | Original | Sham  | Inverted | $\Delta$ pp | 95% CI (pairs)              | Holm $p$ |
|-----------------|----------|-------|----------|-------------|-----------------------------|----------|
| Gemma 3 27B     | 8/32     | 6/32  | 11/32    | +15.6       | [+0.0, +31.2] ( $n = 32$ )  | 1.000    |
| Mistral 3.1 24B | 5/32     | 6/32  | 6/32     | +0.0        | [-9.4, +9.4] ( $n = 32$ )   | 1.000    |
| Qwen2.5-VL 3B   | 1/32     | 4/32  | 5/32     | +3.1        | [-12.5, +18.8] ( $n = 32$ ) | 1.000    |
| GPT-5.6 Sol     | 32/32    | 32/32 | 32/32    | +0.0        | [+0.0, +0.0] ( $n = 32$ )   | 1.000    |

Table 33: Matched replication on the same 32 B groups with separate candidate sets. All four contrasts have Holm  $p = 1$  in the 12-test family. The empirical Sol interval is degenerate at its observed ceiling; it is not equivalence.

### S.1 CROSSED EVIDENCE RESOLVES SPECIFIC OMISSION ACCOUNTS

All 1,408 new requests are recorded: 1,404 valid answers and four completed invalid Mistral outputs (two per crossed family), scored incorrect. There are no unknown or missing outcomes. Through these three studies, before the orientation experiment, this gives 4,096 attempts on 288 base groups, with 4,094 known outcomes (six invalid) and two historical unknowns. Calibration requests are excluded.

Sol answers 123/128 A tasks (96.09%) and 121/128 B tasks (94.53%), with 27/32 and 25/32 entirely correct groups (Table 32). The Hoeffding bounds are  $1.24 \times 10^{-6}$  and  $3.08 \times 10^{-6}$ ; their 12-test Holm values are  $1.49 \times 10^{-5}$  and  $3.39 \times 10^{-5}$ . Thus, under the stated stable-kernel and independence assumptions, its A responses are inconsistent with rules omitting either the initial image or translation instruction, and its B responses with rules omitting axis or order. Complete-group bootstrap accuracy intervals are 92.97–99.22% and 90.63–97.66%; the conservative one-sided bounded-group lower limits are 74.46% and 72.90%. These describe different uncertainty constructions.

The other configurations score 23.44–31.25%, have no entirely correct group, and do not reject the omission nulls. Nonrejection does not establish that a factor was ignored. Descriptively, Sol changes its answer on 58/64 order-switch pairs in B, with both answers correct in 57/64; the labels must change in all 64 pairs. Appendix tables retain every factor, invalid answer, joint-correctness pattern and comparator agreement.

Both radial and relational image-only rules score exactly 32/128 in each crossed family, including after resizing. The literal-translation rule scores 42/128 in A and 32/128 in B; the privileged source executor scores 128/128. The perfect alternative accounts in the first experiment therefore cease to be perfect on these valid crossed questions. This changes what a high score can exclude, without attributing Sol’s internal computation or its earlier matched-task behavior to a particular algorithm.

### S.2 MATCHED REPLICATION REMAINS A BOUNDED NULL

The noncommuting construction reproduces the geometric control: the radial rule gives 32/32 original and sham, then 0/32 inverted, while the relational rule stays at 32/32 throughout. The translation-only rule now gives 6/32, 15/32 and 19/32. No model’s primary intervention–sham contrast is significant (Table 33). Gemma’s five-net-win contrast has six wins, one loss and exact unadjusted  $p = 0.125$ . Sol is correct on all 96 requests; its complete-pair bootstrap is  $[0, 0]$ , while the conservative bounded-difference 95% interval is  $[-48.02, 48.02]$  percentage points. There is no population equivalence claim. The replication establishes a valid second application of the matched control, not a newly significant construction effect. Crossed success and matched robustness remain separate behavioral observations.

## T COUNTERFACTUAL CONSTRUCTION AND VERIFICATION DETAILS

The formal master seed is 202609171127, shuffled task-order seed 202609171128, and bootstrap seed 202609171129. Each of the two families has eight groups per shape. Signed initial-center coordinates have integer magnitudes 10–26; translation-coordinate magnitudes are 8–20. Rectangles have widths 30–38 and heights 20–26; squares have side 26–34; ellipses have dimensions 32–40 by 20–28; hexagons have radius 16–20. Preparation angles come from  $\{\pm 10, \pm 25, \pm 40\}$  degrees. They specify initial shape orientation, not an additional instruction. Source rounding is retained when executing reflection and translation programs.

All task panels use the frame  $[-100, 100]^2$ . A proposal is rejected if an initial/candidate vertex has less than eight units of boundary clearance, resized shape extent is below 12 pixels, crossed candidate centroids are separated by less than 12 units, targets are not distinct/unique, source and independent execution disagree, or required geometry/image disjointness fails. The selected B-matched correct centroid must have radius 24–36 units, a pre-outcome visibility and cue-separation gate. No model correctness enters these decisions. The finite proposal budget is 10,000; thresholds are never relaxed to fill a stratum. The actual 176 proposals yield 64 accepted groups and 112 rejections: 13 fail separation and 99 fail the radius gate. Rejected parameters and partial artifacts remain in the disposition ledger.

Within every B shape, each program is selected twice for the matched block, once with each sham sign. Matched labels are independently balanced at eight per letter; all crossed groups already contain all four labels. B’s crossed and matched blocks reuse 32 groups, not 64. The accepted groups contain 96 unique initial polygon sets and 96 initial-image byte hashes, because A has two initial states per group. These are disjoint from the earlier prospective and isolated calibration states under the recorded exact geometry/image checks; this is not a claim of semantic independence among shape families.

An independent validator checks all 352 tasks, all case parameters, actual prompt/input identities, exact source execution, polygon-set target uniqueness, raw/resized visibility, and matched invariants. Maximum executor discrepancy is zero. Minimum accepted frame clearance is 26.63 units and minimum resized extent is 29 pixels. Across matched triplets, maximum centroid-distance discrepancy is  $1.51 \times 10^{-14}$ , indexed-vertex MSE discrepancy is  $9.10 \times 10^{-13}$ , and equal-movement discrepancy is  $1.42 \times 10^{-14}$ . Correct-panel pixels are identical within every matched triplet before and after resizing. This validates the specified finite geometry; it does not certify equal perceptual difficulty.

Four disjoint calibration groups provide 88 calls, all with valid output format. They check image delivery, parsing and accounting and are excluded from all scientific counts. Formal API runs use two workers, 120-second RealLLM and 180-second Azure request timeouts, and a 3,600-second external wall cap per runner. The local runner uses one GPU and one request at a time. Models never receive file paths, scene IDs, group IDs, allocation metadata or other tasks’ answers. Each call is stateless and order is shuffled. The local model is greedy; deterministic decoding does not imply bitwise kernel reproducibility.

**Inference beyond the displayed task set.** The finite realized counts do not require a sampling model. Hoeffding tests condition on the fixed task set, with the between-group response-randomness and stable-kernel assumptions in Appendix R. For a mean in  $[0, 1]$ , the one-sided 95% lower bound uses  $\bar{Z}_{\text{lower}} - \sqrt{\log(20)/(2N)}$ , clipped at zero. This remains nondegenerate at an empirical ceiling. For matched differences in  $[-1, 1]$ , the two-sided 95% conservative half-width is  $\sqrt{2 \log(40)/N}$ , clipped to the feasible range and kept separate from unknown-outcome identification bounds. These intervals are conservative and not multiplicity-adjusted. An optional binomial all-success bound in the machine-readable outputs is explicitly conditional on a hypothetical iid-group model; balanced generation does not establish that model.

Whole-group and complete-pair percentile bootstraps use 10,000 draws. Their hypothetical resampling interpretation is exploratory, and missing groups/pairs do not become known successes or failures. Required-factor switch counts show whether valid predicted labels change when the answer must change; switching alone is not correctness or mechanism attribution. The 12 primary tests are corrected together, allowing dependence among their valid p-values. The separate earlier

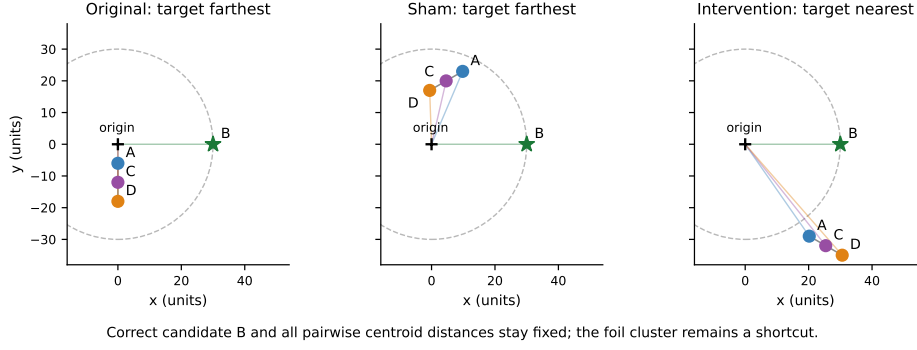


Figure 12: One isolated calibration case, overlaid for visualization; actual tasks use separate panels. Opposite rotations of wrong centroids around correct B (star) preserve pairwise distances and match movement while reversing the radial cue. The correct centroid remains a relational outlier.

four-model matched family is retained as originally frozen, not retroactively pooled with the new confirmatory family.

## U COMPLETE CROSSED AND MATCHED SUMMARIES

| Model           | Family | $K = 0$ | $K = 1$ | $K = 2$ | $K = 3$ | $K = 4$ | Incomplete |
|-----------------|--------|---------|---------|---------|---------|---------|------------|
| Gemma 3 27B     | A      | 2       | 25      | 5       | 0       | 0       | 0          |
| Gemma 3 27B     | B      | 5       | 24      | 3       | 0       | 0       | 0          |
| Mistral 3.1 24B | A      | 3       | 28      | 1       | 0       | 0       | 0          |
| Mistral 3.1 24B | B      | 2       | 27      | 3       | 0       | 0       | 0          |
| Qwen2.5-VL 3B   | A      | 0       | 24      | 8       | 0       | 0       | 0          |
| Qwen2.5-VL 3B   | B      | 2       | 23      | 7       | 0       | 0       | 0          |
| GPT-5.6 Sol     | A      | 0       | 0       | 0       | 5       | 27      | 0          |
| GPT-5.6 Sol     | B      | 0       | 0       | 0       | 7       | 25      | 0          |

Table 34: Number of groups with  $K$  correct answers among four tasks. Completed invalid answers count wrong; no outcomes are unknown in this experiment.

| Model           | Family | Changed factor | Switch/valid | Both correct | U/I pairs |
|-----------------|--------|----------------|--------------|--------------|-----------|
| Gemma 3 27B     | A      | Initial state  | 13/64        | 1/64         | 0/0       |
| Gemma 3 27B     | A      | Translation    | 13/64        | 2/64         | 0/0       |
| Gemma 3 27B     | B      | Axis           | 23/64        | 0/64         | 0/0       |
| Gemma 3 27B     | B      | Order          | 19/64        | 1/64         | 0/0       |
| Mistral 3.1 24B | A      | Initial state  | 10/62        | 0/64         | 0/2       |
| Mistral 3.1 24B | A      | Translation    | 8/62         | 0/64         | 0/2       |
| Mistral 3.1 24B | B      | Axis           | 12/62        | 1/64         | 0/2       |
| Mistral 3.1 24B | B      | Order          | 7/63         | 0/64         | 0/1       |
| Qwen2.5-VL 3B   | A      | Initial state  | 19/64        | 4/64         | 0/0       |
| Qwen2.5-VL 3B   | A      | Translation    | 7/64         | 1/64         | 0/0       |
| Qwen2.5-VL 3B   | B      | Axis           | 16/64        | 2/64         | 0/0       |
| Qwen2.5-VL 3B   | B      | Order          | 15/64        | 1/64         | 0/0       |
| GPT-5.6 Sol     | A      | Initial state  | 61/64        | 59/64        | 0/0       |
| GPT-5.6 Sol     | A      | Translation    | 62/64        | 59/64        | 0/0       |
| GPT-5.6 Sol     | B      | Axis           | 63/64        | 57/64        | 0/0       |
| GPT-5.6 Sol     | B      | Order          | 58/64        | 57/64        | 0/0       |

Table 35: Required-factor changes: 64 pairs clustered in 32 groups per factor. Switches require two valid labels; both-correct counts retain the full 64 pairs. U/I counts pairs with any unknown outcome or a known invalid answer, respectively.

| Rule                                                            | A-crossed | B-crossed | B-original | B-sham | B-inverted |
|-----------------------------------------------------------------|-----------|-----------|------------|--------|------------|
| <i>Comparator correct / planned</i>                             |           |           |            |        |            |
| Origin (raw)                                                    | 32/128    | 32/128    | 32/32      | 32/32  | 0/32       |
| Outlier (raw)                                                   | 32/128    | 32/128    | 32/32      | 32/32  | 32/32      |
| Origin (1344)                                                   | 32/128    | 32/128    | 32/32      | 32/32  | 0/32       |
| Outlier (1344)                                                  | 32/128    | 32/128    | 32/32      | 32/32  | 32/32      |
| Source oracle                                                   | 128/128   | 128/128   | 32/32      | 32/32  | 32/32      |
| Translation only                                                | 42/128    | 32/128    | 6/32       | 15/32  | 19/32      |
| <i>Gemma 3 27B: model agreement / known with comparator</i>     |           |           |            |        |            |
| Origin (raw)                                                    | 49/128    | 34/128    | 8/32       | 6/32   | 4/32       |
| Outlier (raw)                                                   | 29/128    | 38/128    | 8/32       | 6/32   | 11/32      |
| Origin (1344)                                                   | 49/128    | 34/128    | 8/32       | 6/32   | 4/32       |
| Outlier (1344)                                                  | 23/128    | 38/128    | 8/32       | 6/32   | 11/32      |
| Source oracle                                                   | 35/128    | 30/128    | 8/32       | 6/32   | 11/32      |
| Translation only                                                | 24/128    | 24/128    | 8/32       | 11/32  | 6/32       |
| <i>Mistral 3.1 24B: model agreement / known with comparator</i> |           |           |            |        |            |
| Origin (raw)                                                    | 39/128    | 42/128    | 5/32       | 6/32   | 13/32      |
| Outlier (raw)                                                   | 28/128    | 21/128    | 5/32       | 6/32   | 6/32       |
| Origin (1344)                                                   | 39/128    | 42/128    | 5/32       | 6/32   | 13/32      |
| Outlier (1344)                                                  | 20/128    | 21/128    | 5/32       | 6/32   | 6/32       |
| Source oracle                                                   | 30/128    | 33/128    | 5/32       | 6/32   | 6/32       |
| Translation only                                                | 34/128    | 28/128    | 8/32       | 8/32   | 12/32      |
| <i>Qwen2.5-VL 3B: model agreement / known with comparator</i>   |           |           |            |        |            |
| Origin (raw)                                                    | 24/128    | 17/128    | 1/32       | 4/32   | 6/32       |
| Outlier (raw)                                                   | 20/128    | 20/128    | 1/32       | 4/32   | 5/32       |
| Origin (1344)                                                   | 24/128    | 17/128    | 1/32       | 4/32   | 6/32       |
| Outlier (1344)                                                  | 23/128    | 20/128    | 1/32       | 4/32   | 5/32       |
| Source oracle                                                   | 40/128    | 37/128    | 1/32       | 4/32   | 5/32       |
| Translation only                                                | 40/128    | 48/128    | 10/32      | 4/32   | 5/32       |
| <i>GPT-5.6 Sol: model agreement / known with comparator</i>     |           |           |            |        |            |
| Origin (raw)                                                    | 35/128    | 33/128    | 32/32      | 32/32  | 0/32       |
| Outlier (raw)                                                   | 35/128    | 34/128    | 32/32      | 32/32  | 32/32      |
| Origin (1344)                                                   | 35/128    | 33/128    | 32/32      | 32/32  | 0/32       |
| Outlier (1344)                                                  | 33/128    | 34/128    | 32/32      | 32/32  | 32/32      |
| Source oracle                                                   | 123/128   | 121/128   | 32/32      | 32/32  | 32/32      |
| Translation only                                                | 42/128    | 31/128    | 6/32       | 15/32  | 19/32      |

Table 36: All frozen comparator accuracies and model/rule agreement counts. Agreement is descriptive, not an identification of internal strategy.

| Model           | Quantity           | Complete | Full value | Empirical 95% CI | Lower 95% | Bounded 95%    |
|-----------------|--------------------|----------|------------|------------------|-----------|----------------|
| Gemma 3 27B     | A accuracy         | 32/32    | 27.3       | [23.4, 31.2]     | 5.7       | [3.3, 51.4]    |
| Gemma 3 27B     | B accuracy         | 32/32    | 23.4       | [19.5, 27.3]     | 1.8       | [0.0, 47.4]    |
| Gemma 3 27B     | B matched $\Delta$ | 32/32    | +15.6      | [+0.0, +31.2]    | —         | [-32.4, +63.6] |
| Mistral 3.1 24B | A accuracy         | 32/32    | 23.4       | [20.3, 26.6]     | 1.8       | [0.0, 47.4]    |
| Mistral 3.1 24B | B accuracy         | 32/32    | 25.8       | [22.7, 28.9]     | 4.1       | [1.8, 49.8]    |
| Mistral 3.1 24B | B matched $\Delta$ | 32/32    | +0.0       | [-9.4, +9.4]     | —         | [-48.0, +48.0] |
| Qwen2.5-VL 3B   | A accuracy         | 32/32    | 31.2       | [27.3, 35.2]     | 9.6       | [7.2, 55.3]    |
| Qwen2.5-VL 3B   | B accuracy         | 32/32    | 28.9       | [25.0, 33.6]     | 7.3       | [4.9, 52.9]    |
| Qwen2.5-VL 3B   | B matched $\Delta$ | 32/32    | +3.1       | [-12.5, +18.8]   | —         | [-44.9, +51.1] |
| GPT-5.6 Sol     | A accuracy         | 32/32    | 96.1       | [93.0, 99.2]     | 74.5      | [72.1, 100.0]  |
| GPT-5.6 Sol     | B accuracy         | 32/32    | 94.5       | [90.6, 97.7]     | 72.9      | [70.5, 100.0]  |
| GPT-5.6 Sol     | B matched $\Delta$ | 32/32    | +0.0       | [+0.0, +0.0]     | —         | [-48.0, +48.0] |

Table 37: Separate empirical bootstrap intervals, unknown-outcome bounds and conservative independent bounded-group intervals. Accuracy entries are percent; matched differences are percentage points. Lower95 is one-sided; bounded95 is two-sided. None is an equivalence test.

| Population                  | Correct/total | Groups | Accuracy | 95% CI       | Questions with failed panel(s) |
|-----------------------------|---------------|--------|----------|--------------|--------------------------------|
| Text subset (primary)       | 245/326       | 196    | 75.2%    | [68.7, 81.3] | 23                             |
| Visual analogy (new cases)  | 18/47         | 32     | 38.3%    | [22.0, 55.6] | 3                              |
| Text subset (matched pilot) | 43/64         | 64     | 67.2%    | [56.2, 78.1] | 4                              |

Table 38: Frozen pixel rule on different populations. The matched subset is a post-discovery descriptive comparison; transfer uses new STARE rendering case IDs, not a new generator. The final column counts questions with at least one failed panel detection, with their frozen fallback predictions retained; it does not count incorrect answers.

| Model                                  | Direct        | State scaffold | No images     | Options only           |
|----------------------------------------|---------------|----------------|---------------|------------------------|
| <i>STARE 2D text; 64 source groups</i> |               |                |               |                        |
| Gemma 3 27B                            | 53.1% (34/64) | 50.0% (32/64)  | 23.4% (15/64) | 15.6% (10/64)          |
| Mistral Small 3.1 24B                  | 62.5% (40/64) | 57.8% (37/64)  | 26.6% (17/64) | 17.2% (11/64)          |
| Qwen2.5-VL 3B                          | 42.2% (27/64) | —              | —             | 14.1% (9/64)           |
|                                        |               |                |               | 21.9–23.4%             |
| GPT-5.6 Sol                            | 98.4% (63/64) | —              | —             | 14/63 known; 1 unknown |
| Centroid rule <sup>†</sup>             | —             | —              | —             | 67.2% (43/64)          |
| <i>MindEdit L4; 64 source groups</i>   |               |                |               |                        |
| Gemma 3 27B                            | 15.6% (10/64) | 10.9% (7/64)   | 18.8% (12/64) | 10.9% (7/64)           |
| Mistral Small 3.1 24B                  | 9.4% (6/64)   | 10.9% (7/64)   | 9.4% (6/64)   | 23.4% (15/64)          |
| Qwen2.5-VL 3B                          | 12.5% (8/64)  | —              | —             | 15.6% (10/64)          |
| GPT-5.6 Sol                            | 50.0% (32/64) | —              | —             | 17.2% (11/64)          |

Table 39: Correct/planned counts on the same 64 groups per dataset. Unrun conditions are dashes; <sup>†</sup>the pixel rule is a post-discovery descriptive comparison. Sol’s unknown outcome retains full-population bounds.

## V EARLIER MATCHED-CONTROL VISUALIZATION

## W PUBLIC-DATA RESULTS IN FULL

The frozen detector answers 245/326 questions correctly, 75.15% (group-bootstrap 95% interval 68.66–81.25%). All 23 questions with at least one failed panel detection remain in the denominator under the original frozen fallback rule; 9 of these 23 questions are answered correctly. Group-macro accuracy is 71.00%. The hindsight-best constant label reaches only 91/326 (27.91%). A post-result translation split gives 235/249 correct versus 10/77 for other instructions; this is descriptive, not a prespecified mechanism test. The released-population score is exact; its interval describes hypothetical group resampling, not uncertainty about that census.

Among 306 released visual-analogy rows, 259 reuse instruction choice-image bytes. We freeze the remaining 47 rows from 32 absent case IDs before accessing labels. The unchanged detector answers 18/47, 38.30% (22.00–55.56%), retaining three questions with at least one failed panel detection under the same frozen fallback rule (0/3 correct). Its interval includes the uniform four-choice reference; this does not establish transfer beyond that reference.

On matched STARE groups, direct/candidate-only correct counts are 27/9 for Qwen, 34/10 for Gemma and 40/11 for Mistral, each out of 64. All three direct advantages survive the six-contrast Holm correction. The pixel rule reaches 43/64 on these exact questions. Sol reaches 63/64 direct and 14 known correct candidate-only answers plus one unknown, giving a direct advantage bounded by 75.0–76.6 percentage points. A deliberately constructed rule therefore succeeds where generic candidate-only model prompting performs poorly; this neither establishes nor excludes conditional use of a shortcut.

MindEdit gives direct/candidate-only counts 8/10, 10/7 and 6/15 for Qwen, Gemma and Mistral. None is significant in the same six-test family. Mistral’s bootstrap interval excludes zero, but its exact unadjusted test is already nonsignificant ( $p = 0.0636$ , Holm  $p = 0.1907$ ). Sol reaches 32/64 versus 11/64: a 32.8-point advantage (17.2–48.4; two-test Holm  $p = 0.000324$ ). Its later replication reuses the exposed groups and remains exploratory. This photographic task diagnoses information removal; it does not replicate the geometric intervention.

| Model                 | Original | Sham  | Inverted       | $\Delta$ pp [95% CI]        | Holm $p$ |
|-----------------------|----------|-------|----------------|-----------------------------|----------|
| Gemma 3 27B           | 30/96    | 33/96 | 39/96          | +6.2 [-6.2, +18.8]          | 0.835    |
| Mistral Small 3.1 24B | 27/96    | 28/96 | 24/96          | -4.2 [-8.3, -1.0]           | 0.500    |
| Qwen2.5-VL 3B         | 10/96    | 11/96 | 18/96          | +7.3 [-1.0, +15.6]          | 0.500    |
|                       |          |       | 95/96          |                             |          |
|                       |          |       | 1 unknown      | [-1.0, +0.0] bounds         |          |
| GPT-5.6 Sol           | 96/96    | 96/96 | [99.0, 100.0]% | +0.0 [+0.0, +0.0]; $n = 95$ | 1.000    |

Table 40: First matched experiment: 96 paired cases per model. Intervals use complete pairs; Sol also has full-96 bounds. All primary tests are nonsignificant.

## X FIRST MATCHED STUDY: SETTINGS AND FULL RESULTS

The frozen design contains 96 new base cases: 48 with shape and position variation, 48 with position-only variation. Four shapes, three operation chains and four labels are balanced within each stratum. All initial shapes are centered and all chains end in translation. The final centroid alone identifies the answer, so this construction does not require resolving earlier operation order. Geometry, source/independent execution, visibility at original and resized resolution, and correct-panel identity are checked before calls. Of 119 proposals, 23 fail prespecified gates (17 centroid-area agreement, six execution rounding); their records are retained.

The four models receive all 288 tasks in frozen shuffled order. The primary contrast is intervention minus sham accuracy, using exact McNemar tests with Holm correction across four models. Complete base-case pairs are bootstrapped jointly (10,000 draws); unknowns retain full-96 effect bounds. Original/sham and stratum analyses are secondary. These comparisons concern the specified construction effect, not unique attribution to an internal radial computation.

All 1,152 planned requests on 96 new groups are attempted: 1,151 valid answers and one Sol intervention timeout, with no invalid or missing request. No primary intervention–sham contrast survives four-test Holm correction.

Mistral has four losses and zero wins, yielding exact unadjusted  $p = 0.125$  (Holm  $p = 0.5$ ), despite a bootstrap interval excluding zero. Sol answers all 287 known requests correctly; its full-96 intervention–sham effect lies in  $[-1.04, 0]$  points. The empirical complete-pair interval  $[0, 0]$  is a ceiling artifact, not equivalence. Smaller models have low baseline accuracy.

The radial pixel rule changes from 96/96 in original and sham to 0/96 after intervention at both recorded resolutions, verifying the intended cue change. However, a centroid-outlier rule solves 96/96 in every condition. A post-result translation-only rule also solves every case: it selects the centroid nearest the instruction’s final translation, ignoring the initial image and earlier operations. These are demonstrated alternative solvers, not inferred model mechanisms. More samples from this construction cannot eliminate that ambiguity.

## Y ADDITIONAL DIAGNOSTICS AND EXECUTION DETAILS

### Y.1 ANCILLARY STATE-SCAFFOLD DIAGNOSTIC

State prompting changes STARE accuracy from 34/64 to 32/64 for Gemma and 40/64 to 37/64 for Mistral. On MindEdit, the changes are 10/64 to 7/64 and 6/64 to 7/64. All four Holm-adjusted tests have  $p = 1$ . This does not prove that spatial representations are unhelpful: the scaffold is neither trained nor guaranteed correct, and the intervals are imprecise. It does show that these runs cannot support a claim that adding explicit state descriptions reliably improves performance.

### Y.2 EXACT PROMPT CHANGES AND EXECUTION BOUNDARIES

The direct suffix requests a single JSON object with only the key `answer`, containing the uppercase option letter, without explanations. The state suffix additionally asks for `state` and `updates`: an initial spatial description followed by ordered updates with explicit reference frames. It supplies no gold state, demonstrations, or external tool. The no-image prefix states that images are unavailable and asks the model to use only text. Candidate-only prompting states that the original question and

| Dataset       | Model                 | Contrast         | $\Delta$ (pp)               | 95% CI         | Pairs |
|---------------|-----------------------|------------------|-----------------------------|----------------|-------|
| MindEdit L4   | Gemma 3 27B           | State — direct   | -4.7                        | [-10.9, +0.0]  | 64/64 |
| MindEdit L4   | Mistral Small 3.1 24B | State — direct   | +1.6                        | [+0.0, +4.7]   | 64/64 |
| STARE 2D text | Gemma 3 27B           | State — direct   | -3.1                        | [-14.1, +7.8]  | 64/64 |
| STARE 2D text | Mistral Small 3.1 24B | State — direct   | -4.7                        | [-12.5, +3.1]  | 64/64 |
| MindEdit L4   | Qwen2.5-VL 3B         | Direct — options | -3.1                        | [-12.5, +4.7]  | 64/64 |
| MindEdit L4   | Gemma 3 27B           | Direct — options | +4.7                        | [-4.7, +14.1]  | 64/64 |
| MindEdit L4   | Mistral Small 3.1 24B | Direct — options | -14.1                       | [-26.6, -1.6]  | 64/64 |
| STARE 2D text | Qwen2.5-VL 3B         | Direct — options | +28.1                       | [+14.1, +42.2] | 64/64 |
| STARE 2D text | Gemma 3 27B           | Direct — options | +37.5                       | [+23.4, +51.6] | 64/64 |
| STARE 2D text | Mistral Small 3.1 24B | Direct — options | +45.3                       | [+32.8, +57.8] | 64/64 |
| MindEdit L4   | GPT-5.6 Sol           | Direct — options | +32.8                       | [+17.2, +48.4] | 64/64 |
| STARE 2D text | GPT-5.6 Sol           | Direct — options | [+75.0, +76.6] <sup>‡</sup> | [+65.1, +85.7] | 63/64 |

Table 41: Paired differences and unadjusted 95% source-group bootstrap intervals. The text reports multiplicity-adjusted tests. Each comparison has 64 planned pairs. <sup>‡</sup>Sol STARE has 63 complete pairs: its effect bounds retain the unknown pair, while its interval uses only complete pairs.

initial scene are unavailable and requests the most likely candidate from the remaining inputs. Exact UTF-8 prompts and image order are retained in JSONL.

The first API experiment comprises  $128 \times 3 \times 2 = 768$  requests. Its candidate-only extension adds  $128 \times 2 = 256$ . Qwen adds  $128 \times 2 = 256$  local requests. Direct API results are reused for paired comparisons without being counted as new calls. Sol adds another  $128 \times 2 = 256$  attempts, including one infrastructure failure. The total 1,536 attempts reuse just 128 source groups, not 1,536 independent scenes. The prospective intervention adds  $96 \times 3 \times 4 = 1,152$  requests, for 2,688 attempts over 224 model-evaluation source groups through that study. Of those, 2,686 outcomes are known (including two invalid answers scored incorrect), and two remain unknown. The crossed study adds 1,408 attempts on 64 new base groups: 1,404 valid and four completed invalid answers, with no unknown or missing outcomes. Through the affine study, before the orientation experiment, combined totals are 4,096 attempts, 4,094 known outcomes (six invalid) and two unknown, over 288 base groups. Calibration calls are excluded. The earlier provider experiments use two workers, per-request timeouts, an external 30-minute wall cap, and immediate per-response persistence. The local 256-request Qwen job takes approximately 191.12 seconds, including model loading and input preparation; it is inference, not training.

### Y.3 DESCRIPTIVE TRANSFER BREAKDOWN AND DETECTOR EXAMPLES

These analyses were specified after the aggregate pixel results and are descriptive. Of the 47 transfer questions, the fixed rule answers 0/5 with one operation, 8/24 with two, and 10/18 with three. By released difficulty category, counts are 5/15 easy, 7/14 medium, and 6/18 hard. All three questions with a failed detected panel are incorrect; 18/44 with all panels detected are correct. Neither the operation nor difficulty groups are independently matched populations, and these small counts do not isolate the reason for weaker transfer. The transfer label histogram is A:11, B:11, C:13, D:12, giving a hindsight-best constant reference of 13/47. All 47 cases remain in the reported primary transfer denominator.

### Y.4 SECONDARY INTERVENTION SUMMARIES

The two strata differ in candidate centered geometry; both retain the relational centroid shortcut. The following breakdowns are descriptive and use the same 96 base cases, not additional independent evidence. Answer switches can occur without changing correctness, and their frequency is not a mechanism attribution. All generated cases, including 23 rejected proposals among 119 considered, are accounted for before the model outcomes are observed.

### Y.5 REPLAY CHECKS FOR THE EARLIER STUDIES

The retained source audits record STARE semantic checks in `stare-semantic-checks.json` and transfer deduplication in `stare-va-overlap.json`. The earlier paired-study analyzers verify prompt and image hashes, reparse raw responses, retain unknown outcomes, and compute exact counts and paired

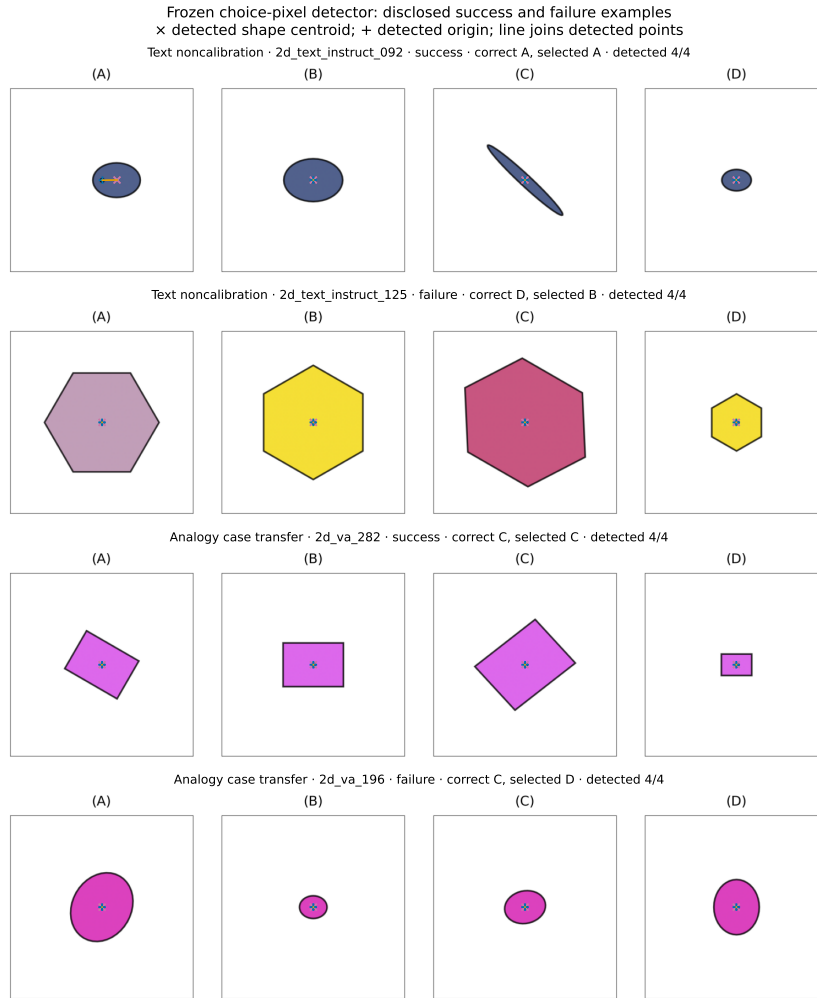


Figure 13: Released-image detector examples. Within each population and correct/incorrect cell, we select the minimum salted hash of the source ID, without a visual-quality filter. Crosses mark detected shape centroids, plus signs mark detected origins, and segments join them. Correct and predicted letters, source IDs, and detected-panel counts are shown. All candidate-image bytes, selected IDs, selection rule, and measurements are retained.

| Model                                | Original | Sham  | Inverted       | $\Delta$ pp [95% CI]        |
|--------------------------------------|----------|-------|----------------|-----------------------------|
| <i>Shape and position; 48 groups</i> |          |       |                |                             |
| Gemma 3 27B                          | 17/48    | 18/48 | 21/48          | +6.2 [-10.4, +22.9]         |
| Mistral Small 3.1 24B                | 15/48    | 16/48 | 14/48          | -4.2 [-10.4, +0.0]          |
| Qwen2.5-VL 3B                        | 5/48     | 11/48 | 12/48          | +2.1 [-10.4, +14.6]         |
| GPT-5.6 Sol                          | 48/48    | 48/48 | 48/48          | +0.0 [+0.0, +0.0]           |
| <i>Position only; 48 groups</i>      |          |       |                |                             |
| Gemma 3 27B                          | 13/48    | 15/48 | 18/48          | +6.2 [-12.5, +25.0]         |
| Mistral Small 3.1 24B                | 12/48    | 12/48 | 10/48          | -4.2 [-10.4, +0.0]          |
| Qwen2.5-VL 3B                        | 5/48     | 0/48  | 6/48           | +12.5 [+4.2, +22.9]         |
| 47/48                                |          |       |                |                             |
| 1 unknown                            |          |       |                |                             |
| [-2.1, +0.0] bounds                  |          |       |                |                             |
| GPT-5.6 Sol                          | 48/48    | 48/48 | [97.9, 100.0]% | +0.0 [+0.0, +0.0]; $n = 47$ |

Table 42: Correct/planned counts in the two prespecified 48-case strata. Unknown outcomes remain explicit. No subgroup hypothesis replaces the primary family.

| Model                 | Switches/valid pairs | C→W | W→C | Known/total | Unknown pairs |
|-----------------------|----------------------|-----|-----|-------------|---------------|
| Gemma 3 27B           | 57/96                | 16  | 22  | 96/96       | 0             |
| Mistral Small 3.1 24B | 7/96                 | 4   | 0   | 96/96       | 0             |
| Qwen2.5-VL 3B         | 43/96                | 5   | 12  | 96/96       | 0             |
| GPT-5.6 Sol           | 0/95                 | 0   | 0   | 95/96       | 1             |

Table 43: Secondary answer-switch accounting on complete condition pairs.

intervals. The pixel-rule protocol fixes thresholds and failure behavior before full evaluation; the separate transfer protocol fixes the 47 eligible image hashes before scoring. The prospective intervention freeze retains generation rejections, response contracts and the primary four-model comparison. Its stored-response replay verifies all 1,152 attempts and preserves the one unknown.

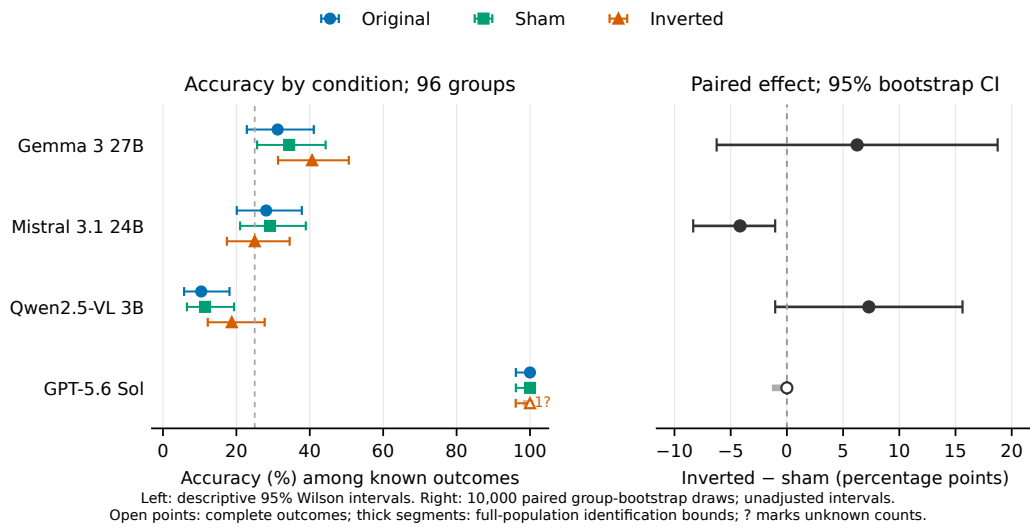


Figure 14: Intervention accuracies and paired contrasts. Individual accuracy intervals are Wilson intervals; the effect panel uses paired bootstrap intervals. Sol’s open intervention marker and full-population effect bounds disclose its one unknown result. Its zero-width complete-pair interval reflects the observed ceiling.