REWARDLENS: MEASURING THE BEHAVIORAL RESOLUTION OF STATIC MULTIMODAL JUDGE SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Static preference accuracy is a common way to compare multimodal judges. A tie shows that two judges were equally accurate on the evaluated examples, but says little about how they respond when decision-relevant visual evidence changes. We call the extent to which a static score constrains such intervention behavior its *behavioral resolution*. RewardLens measures this property with matched visual triplets that hold the question and candidate responses fixed: relevant edits reverse the programmed gold preference, while matched irrelevant edits preserve it. We quantify the two responses with Relevant Adaptation (RA) and Irrelevant Invariance (II). On an independent same-domain CLEVR static evaluation, the same five judges score 100% on both Attribute and Spatial, yet their RA spans are 0 and 43.5 percentage points. In Count, Qwen and Skywork score 100.0% and 99.5% statically but differ by 45.73 points in RA on shared correct bases while remaining nearly perfectly invariant to irrelevant edits. Across all 42 frozen one-point same-domain matches, 16 have zero RA gap and 16 have gaps of at least 10 points. Static accuracy summarizes correctness on fixed examples. When responses to changing evidence matter, that summary should be paired with controlled intervention measurements.

1 INTRODUCTION

Static preference accuracy is a natural way to compare multimodal judges: on a fixed collection of image–question–response pairs, how often does each judge select the correct response? That number records performance on the evaluated items, but compresses each judge's response function. It leaves open whether tied judges will respond alike when the visual evidence changes while the question and candidate responses stay fixed.

We study this missing information as a *behavioral resolution* problem: how tightly does a static evaluation constrain intervention behavior among judges that it assigns the same score? A comparison across Attribute and Spatial illustrates the issue. The same five judges—Idefics3, InternVL, LLaVA, Qwen, and Skywork—all score 100% on independent CLEVR Attribute items and on independent CLEVR Spatial items. Yet their Relevant Adaptation (RA) spans are 0 pp on Attribute and 43.5 pp on Spatial, with all five correct on all 200 audit bases in each factor (Fig. 1). Thus, the same judge set can form a behaviorally tight exact-score class in one factor and a broad one in another.

RewardLens makes this distinction measurable using controlled CLEVR triplets. The question and candidate responses remain fixed. A relevant edit changes the visual fact that determines the programmed preference, whereas a matched irrelevant edit applies the same class of low-level operation without changing that preference. Conditional on a correct base decision, RA measures correctness after the relevant edit and Irrelevant Invariance (II) after the irrelevant edit. Static matching uses separate evaluation items, so intervention behavior is never used to form score-matched comparisons.

Perturbation audits measure how a model or judge responds to a transformation (Lee et al., 2026; Park et al., 2026; Zhu et al., 2026). RewardLens addresses a different comparison: once a static benchmark has grouped judges by score, it measures how similar their responses remain under controlled evidence changes. An exact tie can have a narrow or wide intervention profile, and a near-tie can conceal selective adaptation differences that are invisible to the scalar.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

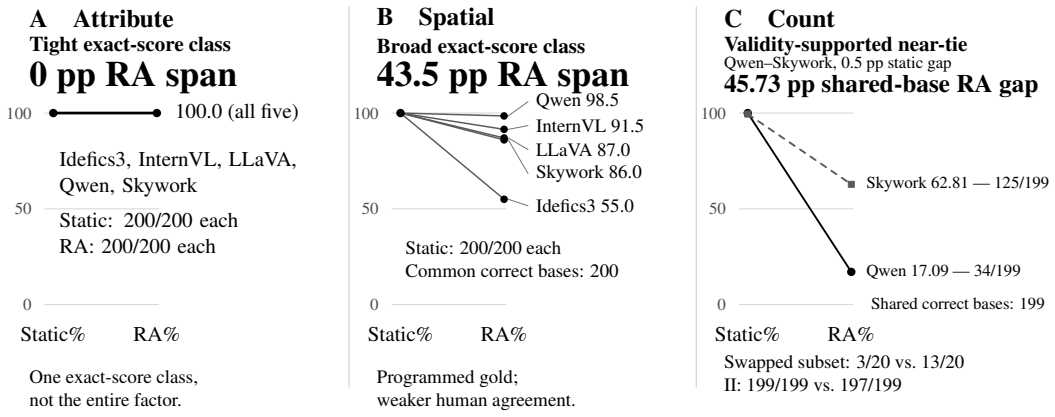


Figure 1: **The same static-score equivalence can have very different behavioral resolution.** (A–B) The same five judges score 200/200 on the independent Attribute and Spatial static sets, but their shared-base RA spans are 0 and 43.5 pp. (C) Qwen–Skywork is a 0.5 pp Count near-tie with a 45.73 pp shared-base RA gap while II remains near-perfect. Spatial uses programmed gold and has weaker absolute human/gold agreement. The order diagnostic uses 20 triplets; lines connect summary values, not paired items.

Across eight judges and four visual factors, the answer is heterogeneous. The frozen one-point same-domain census contains 42 pairs: 16 have zero RA gap, 16 have gaps of at least 10 pp, and the median gap is 2 pp. Count provides the strongest validity-supported case: Qwen and Skywork score 100.0% and 99.5% on independent static items, yet differ by **45.73 pp** in RA on 199 shared correct audit bases while remaining nearly perfectly invariant to irrelevant edits. The contrast also persists on the available candidate-order subset and in a post-result human/gold-concordant stratum. These diagnostics cover restricted subsets and show that the Count contrast persists under both checks.

RewardLens targets observable behavior under finite-sample score matching. Universal judge rankings, latent competence, and internal mechanisms are separate questions. A blinded two-person perceptual audit further shows high self-relative flip/preserve agreement but weaker absolute agreement with programmed gold for Presence and Spatial, so we interpret those factors as contract-specific behavioral evidence and retain perceptual ambiguity as a limitation.

Contributions. We (i) formulate behavioral resolution as a meta-evaluation property of score-based judge comparisons; (ii) introduce a paired relevant/irrelevant intervention audit that separates selective adaptation from generic response instability; and (iii) characterize both tight and broad comparison regimes across eight judges and four factors, with complete matched-pair reporting and common-support, candidate-order, and human validity diagnostics.

2 Related Work

Judge evaluation and multimodal reward models. LLM-as-a-judge work evaluates agreement with human preferences while documenting position, verbosity, and other judgment biases (Zheng et al., 2023). JudgeBench and JETTS extend evaluation to challenging response pairs and test-time selection (Tan et al., 2025; Zhou et al., 2025); multimodal benchmarks evaluate scoring, pairwise comparison, and ranking across image-conditioned tasks (Chen et al., 2024; 2025; Hu et al., 2026). Specialized reward models such as Omni-Reward, R1-Reward, and BaseReward provide additional judges for these settings (Jin et al., 2026; Zhang et al., 2026a;b). These benchmarks evaluate judges on fixed examples. RewardLens uses the resulting static scores to define comparison groups, then measures the intervention behavior that varies within each group.

Controlled behavioral tests. CheckList separates aggregate accuracy from capability-targeted behavior (Ribeiro et al., 2020), while shortcut learning explains why benchmark success can fail under controlled changes (Geirhos et al., 2020). In vision–language evaluation, VQA v2 uses complementary images (Goyal et al., 2017); VALSE, Winoground, CREPE, and SugarCrepe use validated

foils or contrastive examples to expose compositional failures (Parcalabescu et al., 2022; Thrush et al., 2022; Ma et al., 2023; Hsieh et al., 2023). Visual programming provides another route to verifiable counterfactual examples (Gao et al., 2025). The common thread is the use of controlled contrasts. RewardLens applies the same idea to score-matched judges rather than to answer models or representations.

Perturbation audits of multimodal judges. Reward models can exploit text-only correlations, and image manipulations can inflate visual-judge scores (Li et al., 2025; Hwang et al., 2025). MM-JudgeBias measures score deviation under semantic disruption and conformity under semantics-preserving changes, with targeted position- and verbosity-bias analyses (Lee et al., 2026). VisualFLIP holds the question fixed, changes task-critical visual evidence so the gold answer flips, and evaluates competence-conditioned response changes (Zhu et al., 2026). Perceptual Judgment Bias studies conflicts between visual evidence and plausible textual cues, while Multi-Crit studies criterion following (Park et al., 2026; Xiong et al., 2026). RewardLens builds on these perturbation designs and uses them for a meta-evaluation question: after a static score has grouped judges together, how similarly do they respond to controlled evidence changes? The novelty lies in conditioning this comparison on an independently measured static score.

3 Problem Formulation

Let a multimodal judge f receive $x = (I, q, r_1, r_2)$ and return a binary preference with gold label y. For visual factor c, its measured static accuracy on an independent set is

$$\widehat{A}_{S,c}(f) = \frac{1}{n_c} \sum_{k=1}^{n_c} \mathbf{1}\{f(x_k) = y_k\}. \tag{1}$$

Throughout, equality refers to these finite-sample estimates.

3.1 Controlled Triplets and Behavioral Measures

An audit triplet contains $x_B = (I_B, q, r_1, r_2)$, $x_R = (I_R, q, r_1, r_2)$, and $x_I = (I_I, q, r_1, r_2)$. The question and candidate responses remain fixed. The programmatic scene contract requires $y_R \neq y_B$ for the relevant edit and $y_I = y_B$ for the irrelevant edit; Section 6.3 separately tests human agreement with these labels on a frozen subset. Across m_c triplets, RewardLens conditions on a correct base decision and estimates

$$\widehat{\mathrm{RA}}_c(f) = \frac{\sum_k \mathbf{1}\{f(x_{B,k}) = y_{B,k} \wedge f(x_{R,k}) = y_{R,k}\}}{\sum_k \mathbf{1}\{f(x_{B,k}) = y_{B,k}\}}, \tag{2}$$

$$\widehat{\mathrm{II}}_c(f) = \frac{\sum_k \mathbf{1}\{f(x_{B,k}) = y_{B,k} \wedge f(x_{I,k}) = y_{I,k}\}}{\sum_k \mathbf{1}\{f(x_{B,k}) = y_{B,k}\}}. \tag{3}$$

The denominator is judge-specific. We report audit-base accuracy $\widehat{A}_{B,c}$ separately and also compare judges on the intersection of their base-correct sets (Section 7). Every reported model–factor cell has nonzero base-correct support.

With a binary output, a base-correct judge succeeds on the relevant branch exactly when it changes its prediction. Hence

$$1 - \mathrm{RA}_c(f) = P(f(x_R) = f(x_B) \mid f(x_B) = y_B). \tag{4}$$

Prediction retention is a failure on the relevant branch but the desired behavior on the irrelevant branch. RA and II therefore separate selective adaptation from indiscriminate instability.

3.2 Behavioral Resolution of Static Scores

For the evaluated panel $\mathcal{J}$, an exact measured static score s induces the score-equivalence class

$$\mathcal{E}_c(s) = \{f \in \mathcal{J} : \widehat{A}_{S,c}(f) = s\}. \tag{5}$$

Behavioral resolution describes how tightly the static evaluation constrains intervention behavior within the class it induces. For RA we use the within-class span

$$\mathrm{Span}_{\mathrm{RA}}(\mathcal{E}_c(s)) = \max_{f \in \mathcal{E}_c(s)} \widehat{\mathrm{RA}}_c(f) - \min_{f \in \mathcal{E}_c(s)} \widehat{\mathrm{RA}}_c(f). \tag{6}$$

Smaller spans indicate tighter behavioral alignment within the measured class; larger spans indicate greater RA variation among its members. We use the span only to summarize observed RA values for each non-singleton class. Itemwise agreement and population-level identifiability require separate analyses.

For near-ties, judges f and g are matched within factor c at tolerance ϵ when $|\widehat{A}_{S,c}(f) - \widehat{A}_{S,c}(g)| \leq \epsilon$. Because this pairwise rule is not transitive for $\epsilon > 0$, it is distinct from an exact-score class. We report residual gaps

$$\Delta_{\mathrm{RA},c} = |\widehat{\mathrm{RA}}_c(f) - \widehat{\mathrm{RA}}_c(g)|, \qquad \Delta_{\mathrm{II},c} = |\widehat{\mathrm{II}}_c(f) - \widehat{\mathrm{II}}_c(g)|. \tag{7}$$

Because the static and audit items are disjoint, two judges can tie statically while responding differently on the audit branches. We examine the observed structure of those differences: whether exact-score classes are tight or broad, whether near-ties contain large selective-response gaps, and whether the pattern changes across factors.

3.3 Analysis Units and Interpretation

We use an independent natural-image static set and a separately generated CLEVR static control. The intervention triplets are disjoint from both sets. Accordingly, $\widehat{A}_{S,c}$ measures independent static accuracy, whereas RA/II characterize audit behavior. The CLEVR control places matching and intervention in the same visual domain. Each tie retains its finite-sample interpretation. Matching produces a descriptive within-factor census over eight judges. Qualifying pairs can share a judge and therefore are not independent observations. For the original natural-image pairs, paired-bootstrap intervals jointly resample common static-item and source-triplet identifiers. The same-domain analysis is a census of every pair selected by the frozen matching rule; because pairs can share judges, we report their gaps descriptively rather than attach post-match confidence intervals to the collection. Appendix Table 2 summarizes the comparison regimes and associated diagnostics.

4 RewardLens Audit Construction

RewardLens-CLEVR instantiates factor-specific intervention contracts in CLEVR-style scenes (Johnson et al., 2017). Within each triplet, the question, candidate responses, candidate order, camera, lighting, and render seed remain fixed; only the image changes. The relevant and irrelevant branches use the same class of low-level operation but differ in decision relevance.

For **Count**, the branches add a target-color versus non-target-color object; for **Attribute**, they recolor the queried object versus a matched non-target; for **Presence**, they remove the queried color–shape conjunction versus a matched non-target; and for **Spatial**, they move the queried referent versus a matched distractor. By construction, $y_R \neq y_B$ for relevant edits and $y_I = y_B$ for irrelevant edits.

Factor-specific post-render validators recompute the realized factor from the rendered scene state and enforce the programmed contract. Human perceptual agreement is evaluated separately in Section 5.3. Full contracts, validation rules, and edit diagnostics appear in Appendix C.3.

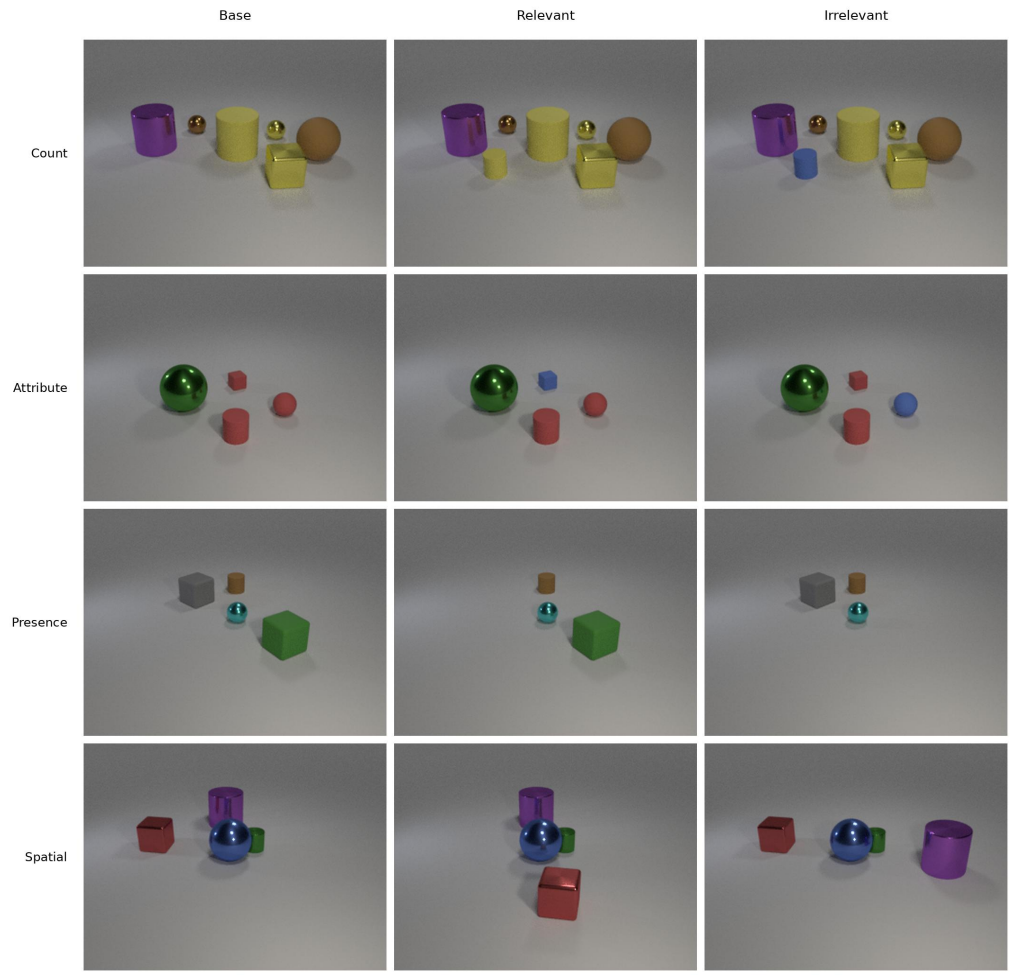


Figure 2: **Representative frozen RewardLens-CLEVR intervention triplets.** Rows show Count, Attribute, Presence, and Spatial; columns show Base, Relevant, and Irrelevant images. Within each triplet, the question and candidate responses stay fixed. Relevant edits reverse the programmed gold preference, whereas matched irrelevant edits preserve it.

5 Experimental Setup

5.1 Evaluation Environments

The original independent natural-image static evaluation has 200 frozen items per factor (800 total). Count uses TallyQA (Acharya et al., 2019); Attribute, Presence, and Spatial use GQA (Hudson & Manning, 2019). Each item pairs a gold response with a structured alternative, with candidate order shuffled deterministically. The separate RewardLens-CLEVR audit has 200 programmatically validated triplets per factor (800 triplets; 2,400 rendered images). Static and audit examples are disjoint, and all judges receive the same items. Natural-image static estimates use 200 reported outputs, except for Phi-3.5-Vision on Spatial ($n = 198$), whose estimate uses the available outputs. The independent same-domain control follows prospectively frozen CLEVR static protocol v2.1. It comprises 800 new base-only scenes and images, 200 per factor, with exactly 100 gold A and 100 gold B positions per factor. The manifest, scene and image assets, validation reports, and dataset freeze receipt preceded inference. Each of the eight historical judges produced 800 judgments ($6{,}400$ total), with no parse errors. Checks possible with the available scene signatures and intervention metadata passed. The historical audit lacks some overlap fields, so the report records the checks

supported by the available metadata. The control uses a new CLEVR static sample independent of the audit bases.

5.2 Multimodal Judges and Inference

We evaluate eight multimodal judges from different model families and training objectives: Qwen3-VL-4B, Gemma-3-4B, Molmo-7B-D, Skywork-VL-Reward-7B, Idefics3-8B, Phi-3.5-Vision, InternVL3-8B, and LLaVA-OneVision-7B. A common deterministic prompt requests exactly A or B. Model-specific adapters parse the binary decision; Skywork instead uses its reward head. Checkpoints, adapters, the prompt, and decoding settings appear in Appendix C.1. Candidate order is fixed across the three states of each triplet, leaving the visual input as the manipulated variable. RA and II use the judge-specific base-correct denominator in Eqs. 2–3. On a frozen subset of 80 audit triplets (20 per factor), the candidate-order diagnostic swaps only A/B content. It retains the original-order judgments and collects three swapped states per triplet and model (1,920 new judgments). Printed decisions are mapped back to candidate content; results are reported separately by orientation, including common base-correct support.

5.3 Blinded Human Perceptual Audit

Two human annotators separately judged the same frozen 320 triplets (80 per factor; 960 images). For the base and two edited images, each chose A, B, or *unable to determine*; relevant/irrelevant roles were concealed and balanced. They also marked judgeability and obvious visual ambiguity, without seeing gold, models, or each other's responses. Both raw submissions were hashed before role unblinding. A1 used one triplet per page and A2 used ten per scroll page, with identical response fields. Analyses are post-freeze descriptive and retain every item and disagreement.

5.4 Static Matching and Analysis

The frozen analysis plan specifies within-factor matching bands of 1, 2, and 3 percentage points. The same bands are applied separately to the original natural-image and new same-domain static scores, as specified before either set of results was examined. The primary rule is

$$|\widehat{A}_{S,c}(f) - \widehat{A}_{S,c}(g)| \leq 0.01. \tag{8}$$

We report every qualifying pair. Because pairs can share a judge, pair counts form a descriptive census. We also report paired-bootstrap intervals and common-base-correct comparisons. The later exact-class summaries group judges by equality of unrounded static fractions. Appendices C.4 and C.5 provide provenance, data receipts, and commands.

6 Results: Behavioral Resolution, Structure, and Validity

6.1 RQ1: Behavioral Resolution under Same-Domain Score Matching

Applying the frozen 1 pp rule to the independent CLEVR static scores yields 42 within-factor matches (Table 1). Their intervention behavior is heterogeneous: 16 pairs have zero RA gap, 16 have absolute gaps of at least 10 pp, and the median gap is 2.00 pp. Pairs may share judges and are descriptive, not independent replicates.

Table 1: **Same-domain matched-pair census under the frozen 1 pp static rule.** The 42 pairs include both zero-gap and large-gap regimes. Individual pairs appear in Appendix B.1.

Factor	Pairs	ΔRA$=0$	ΔRA≥ 10 pp	Median (pp)	Max (pp)
Count	3	0	2	44.81	45.81
Attribute	13	10	2	0.00	14.02
Presence	15	6	5	2.00	39.90
Spatial	11	0	7	12.50	43.50
All	**42**	**16**	**16**	**2.00**	**45.81**

The factor-level structure differs markedly. Attribute is dominated by zero aggregate RA gaps (10 of 13), whereas Spatial has no zero-gap pair and a 12.50 pp median gap. Count has only three qualifying pairs, two with large gaps; Presence mixes zero and large gaps. Static proximity therefore does not reveal which behavioral regime applies.

Relation to the original natural-image matches. The original natural-image analysis contains seven frozen one-point pairs, but only two remain within 1 pp on CLEVR. It is therefore complementary cross-domain evidence rather than a same-domain replication; all original pairs and intervals remain in Table 4.

6.2 RQ2: The Same Judge Set Can Form a Tight Exact-Score Class in One Factor and a Broad One in Another

The same five judges form the complete exact 100% classes in both Attribute and Spatial: Idefics3, InternVL, LLaVA, Qwen, and Skywork. Their Attribute RA is 200/200 for every member, a **0 pp** span. Their Spatial RA is respectively 110/200, 183/200, 174/200, 197/200, and 172/200, a **43.5 pp** span (Fig. 1). All five judges are correct on all 200 audit bases in each factor, and II is 200/200 throughout, so the contrast does not arise from unequal base-correct support or generic sensitivity to the irrelevant branch.

This is one shared judge set, not a factor-wide law. The other exact Attribute class, Gemma–Molmo at 96.5% static accuracy, has a 3.78 pp RA span. All six non-singleton exact-score classes are reported in Table 12, including small-span classes. Twelve of 32 static model–factor cells are at 100%, so the finite static set has limited discriminative resolution at the ceiling. The intervention audit shows that such a tie can be either behaviorally tight or broad; it does not imply equal ability on harder unseen items. Spatial remains relative to programmed gold, for which the human audit shows weaker absolute agreement.

6.3 RQ3: Count as a Validity-Supported Near-Tie

Count gives the strongest validity-supported near-tie. Qwen and Skywork score 100.0% and 99.5% on independent static Count items, but their full-audit RA is $34/200$ (17.00%) versus $125/199$ (62.81%). On their 199 shared correct bases, Qwen is $34/199$ and Skywork $125/199$, a **45.73 pp** gap. II is $200/200$ versus $197/199$. The contrast is therefore selective: the judges differ sharply after the gold-changing edit while remaining nearly invariant to the gold-preserving edit.

The joint transition partition makes the behavior concrete (Table 13). Of Qwen's 200 correct bases, 34 show a relevant update with irrelevant preservation and 166 retain the base choice on both branches. Of Skywork's 199, the corresponding counts are 123 and 74; two additional cases change on both branches. These are exhaustive descriptive partitions, not model mechanisms or new primary scores. The complete 32-cell RA/II ledger appears in Fig. 3.

On the frozen 20-triplet order subset, both judges are base-correct in both orientations. Their RA is $1/20$ versus $13/20$ originally and $3/20$ versus $13/20$ after swapping candidate content. Candidate position therefore does not account for this contrast on that subset, although other comparisons are order-sensitive (Section 7.4).

Human transitions and programmed-gold agreement. Relative to each annotator's own base choice, both humans show the expected relevant-flip/irrelevant-preserve pattern on 280/320 triplets, including 76/80 Spatial triplets. This self-relative agreement does not establish that the programmed label is perceptually unambiguous. Both annotators select all three programmed labels on 195/320 triplets: 58/80 Count, 60/80 Attribute, 41/80 Presence, and 36/80 Spatial (Table 14). The scene-level recheck reproduces all 960 sampled programmatic labels, while human/gold disagreement is concentrated in A=no/B=yes presentations for Presence and Spatial (Table 17).

A post-result Count check crosses human/gold agreement with the machine contrast. Among 58 triplets where both humans match all three labels, Qwen and Skywork share 57 correct bases and have RA $12/57$ versus $35/57$, a 40.35 pp gap. The complementary 22 human-audited triplets give $3/22$ versus $12/22$ (40.91 pp). The contrast is therefore not confined to human/gold-disagreement items, but these outcome-selected strata cannot replace the full audit. Only five concordant triplets

also belong to the order subset, so the controls do not jointly validate all 200 Count triplets (Table 18).

7 Robustness and Validity Checks

7.1 Shared-Base-Correct Comparisons

On the 186 triplets whose base images both Phi and LLaVA judge correctly, their Attribute RA remains $84.95\%/100.00\%$. Across all seven one-point matched pairs, restriction to shared base-correct support leaves the descriptive median absolute RA gap at 15.05 pp. This common-support diagnostic complements, rather than replaces, judge-conditional RA. Individual comparisons can nevertheless change. On Count, Molmo and Phi have judge-conditional RA of 18.63% and 20.00%, but 17.00% and 4.00% on their 100 shared base-correct triplets. The signed Molmo-minus-Phi contrast reverses from -1.37 to $+13.00$ pp. A stable median therefore does not imply stable pairwise rankings (Table 7).

7.2 Audit-Base Accuracy

Molmo–Idefics3 Presence has natural-image static accuracies of $79.5\%/79.0\%$ and audit-base accuracies of $99.5\%/100.0\%$. Its 25.63 pp RA gap persists on 199 shared correct bases, as does Molmo–InternVL's. These original comparisons exclude an explanation based solely on one judge failing most bases, not differences specific to edited scenes or human/gold disagreement. Neither pair remains a 1 pp same-domain match. Table 5 gives all conditional denominators.

7.3 Matching-Threshold Sensitivity

Natural-image median absolute RA gaps are $15.05/15.05/15.14$ pp at the frozen 1/2/3 pp bands. The same bands yield 42/44/54 CLEVR pairs, whose 1 pp median is 2.00 pp. Different selected pairs prevent a domain-effect interpretation of the medians. Complete results appear in Table 8 and Appendix B.1.

7.4 Candidate-Order Sensitivity

The frozen 80-triplet order diagnostic collected $1{,}920$ swapped-order judgments, 240 per model. Among the seven original natural-image matched-pair comparisons, three are descriptively stable, three partially sensitive, and one strongly sensitive (Appendix B.4). The Phi–LLaVA Attribute RA contrast persists on its 20-triplet subset: $89.5\%/100\%$ in the original order and $87.5\%/100\%$ after swapping. On the 16 triplets where Phi is base-correct in both orientations, its RA is 87.5% in each; LLaVA remains at 100%. For the same-domain Qwen–Skywork Count comparison in Section 6.3, semantic candidate-content consistency is $58/60$ versus $60/60$. Both judges share all 20 correct bases in each order. In contrast, Molmo Count RA changes from $4/11$ to $0/9$, with no triplet base-correct in both orders; the Molmo–Phi Count contrast reverses direction. We neither label an average across orientations "position-bias-free" nor extrapolate this 20-triplet diagnostic to all 200 audit triplets. Additional missing-output and edit-magnitude sensitivity checks appear in Appendix A.3; neither establishes internal grounding.

8 Discussion

What a static tie does and does not resolve. Matched pairs range from identical RA to large gaps. A static tie can therefore correspond to either regime. Attribute and Spatial make this visible with the same five judges: both factors produce an exact 100% static class, yet one has a 0 pp RA span and the other a 43.5 pp span. In this sense, behavioral resolution is a property of a particular evaluation-induced comparison and can vary across panels, benchmarks, and factors.

Why the irrelevant branch matters. A large RA gap could otherwise reflect generic sensitivity to any image edit. Count shows a more selective pattern. Qwen predominantly retains its base choice on both branches, whereas Skywork more often updates after the relevant edit while preserving its

choice after the irrelevant edit; II remains near-perfect for both. The transition partitions summarize this response pattern. The human audit separately shows high self-relative transition agreement alongside weaker absolute agreement with programmed gold for Presence and Spatial.

Evaluation implication. The practical question is how much a reported accuracy tells us about responses to changing evidence. Static accuracy remains useful for fixed-example correctness. But when judges are used for model comparison, preference filtering, or reward-based selection, a score tie can leave unresolved whether they respond similarly to changing decision-relevant evidence. A controlled intervention profile adds that information to the scalar score. Downstream effects and judge selection require separate evaluation.

Limitations and scope. The findings are specific to the evaluated model panel, finite static sets, visual factors, and intervention families. The CLEVR control has substantial ceiling effects; a harder static set may break some ties. The audit uses constructed scenes and programmed contracts; candidate-order sensitivity is heterogeneous, joint human/order support is small, and weaker human/gold agreement qualifies Presence and Spatial. We therefore interpret the study as a behavioral audit of this panel and these interventions. Internal mechanisms, deployment effects, external predictive validity, training benefits, and universal judge rankings remain open questions.

9 Conclusion

A static multimodal-judge score summarizes correctness on a fixed sample; RewardLens examines whether score-matched judges also respond similarly when the evidence changes. In the same panel, an exact 100% class is behaviorally tight on Attribute (0 pp RA span) but broad on Spatial (43.5 pp), while the Qwen–Skywork Count near-tie shows a 45.73 pp shared-base RA gap with near-perfect irrelevant invariance. Across all 42 same-domain one-point matches, both zero and large gaps occur. Controlled interventions provide a complementary view of how score-matched judges respond when the evidence changes.

AI USE STATEMENT

Generative AI tools supported brainstorming and hypothesis refinement; critiques of the experimental design; implementation and coding; checks of data analysis and result interpretation; manuscript organization and language editing; LaTeX formatting; and reference discovery. The authors made the research decisions, ran or supervised all experiments, verified the numerical results against the retained artifacts, checked the scientific claims, and checked each citation against its original source. They take full responsibility for the manuscript.

ETHICS STATEMENT

RewardLens evaluates publicly available model checkpoints using existing benchmark material and procedurally rendered scenes. A limited perceptual validity audit additionally uses two human annotators, reported only as A1 and A2, who judged benchmark images using the response fields described in Section 5.3. The analysis uses only item-level benchmark judgments and no demographic or sensitive personal attributes. Human responses assess agreement with the programmed evaluation contract. Model results are reported as benchmark behaviors.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 specify triplet construction, the two independent static sets, sample sizes, models, and evaluation rules. Appendices A, B, and C provide the complete primary matched-pair censuses, orientation and human-audit diagnostics, protocol boundaries, and retained freeze and completion receipts.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019. doi: 10.1609/aaai.v33i01.33018076.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595, 2024.

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria Huang, Canyu Chen, Haoqin Tu, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? In *Advances in Neural Information Processing Systems*, 2025.

Minghe Gao, Zhongqi Yue, Wenjie Yan, Yihao Hu, Wei Ji, Siliang Tang, Jun Xiao, Tat-Seng Chua, Yueting Zhuang, and Juncheng Li. Counterfactual evolution of multimodal datasets via visual programming. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2739. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/75c2ec5f98d7b2f50ad68033d2c07086-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

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. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the LVLM judges: Visual biases in LVLM-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23186–23205, 2025. doi: 10.18653/v1/2025.emnlp-main.1182. URL https://aclanthology.org/2025.emnlp-main.1182/.

Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Omni-Reward: Towards generalist omni-modal reward modeling with free-form preferences. In *International Conference on Learning Representations*, 2026.

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

Sua Lee, Sanghee Park, and Jinbae Im. Mm-judgebias: A benchmark for evaluating compositional biases in mllm-as-a-judge. *arXiv preprint arXiv:2604.18164*, 2026.

Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36281–36297, 2025.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ma_CREPE_Can_Vision-Language_Foundation_Models_Reason_Compositionally_CVPR_2023_paper.html.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8253–8280, 2022. doi: 10.18653/v1/2022.acl-long.567. URL https://aclanthology.org/2022.acl-long.567/.

Seojeong Park, Jiho Choi, Junyong Kang, Seonho Lee, Jaeyo Shin, and Hyunjung Shim. Mitigating perceptual judgment bias in multimodal llm-as-a-judge via perceptual perturbation and reward modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.02578.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhu Chen, and Heng Huang. Multi-Crit: Benchmarking multimodal judges on pluralistic criteria-following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8641–8652, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xiong_Multi-Crit_Benchmarking_Multimodal_Judges_on_Pluralistic_Criteria-Following_CVPR_2026_paper.html.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, Haojie Ding, Jiankang Chen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Guorui Zhou, and Liang Wang. R1-Reward: Training multimodal reward model through stable reinforcement learning. In *International Conference on Learning Representations*, 2026a.

Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Kai Wu, Bo Cui, Xu Wang, Jianfei Pan, Haotian Wang, Zhang Zhang, and Liang Wang. BaseReward: A strong baseline for multimodal reward model. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=EuN5iszF0a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The JETTS benchmark of LLM-as-judges as test-time scaling evaluators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79436–79471, 2025.

Didi Zhu, Changrui Chen, Stefanos Zafeiriou, and Jiankang Deng. Visualflip: Do predictions depend on task-critical visual evidence in multimodal reasoning? *arXiv preprint arXiv:2606.07872*, 2026.

A Additional Results

This appendix gives the complete factor-level result ledger and the robustness summaries cited in the main text. Percentages are rounded to two decimals. A_S refers to the original independent natural-image static set; A_B is unconditioned accuracy on RewardLens-CLEVR base images. RA and II condition on a correct base decision.

A.1 RewardLens Comparison Regimes

Table 2: Comparison regimes in RewardLens. Exact score equality induces score-equivalence classes summarized by within-class RA span, whereas approximate matching produces pairwise RA/II gaps. The remaining rows are diagnostics rather than alternative primary metrics.

Regime	Selection, analysis unit, and summary	Role
Exact-score equivalence	$\widehat{A}_{S,c}(f) = s$; score-equivalence class $\mathcal{E}_c(s)$; $\mathrm{Span}_{\mathrm{RA}}(\mathcal{E}_c(s))$	Behavioral resolution: tight vs. broad exact-score class
Near-score matching	$\lvert\widehat{A}_{S,c}(f) - \widehat{A}_{S,c}(g)\rvert \le \epsilon$; judge pair; ΔRA, ΔII	Residual behavioral gap under the frozen matching rule
Shared-base diagnostic	Both judges correct on the same audit bases; judge pair on common support; shared-support RA / II	Conditioning sensitivity
Candidate-order diagnostic	Frozen triplets with A/B content swapped; judge or pair; orientation-specific RA / II	Candidate-order sensitivity
Human-validity diagnostic	Frozen human-audited subset; triplet subset; transition and programmed-gold agreement	Perceptual interpretation of the intervention contract

	Relevant Adaptation (RA)				Irrelevant Invariance (II)			
Model	Count	Attribute	Presence	Spatial	Count	Attribute	Presence	Spatial
Qwen	0.17	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Gemma	0.32	0.95	0.60	0.00	0.88	0.97	1.00	0.98
Molmo	0.19	0.99	0.74	0.40	0.97	0.99	0.99	0.99
Skywork	0.63	1.00	1.00	0.86	0.99	1.00	0.98	1.00
Idefics3	0.71	1.00	1.00	0.55	0.96	1.00	0.99	1.00
Phi	0.20	0.85	0.74	0.47	0.95	0.89	0.97	0.98
LLaVA	0.42	1.00	0.98	0.87	0.99	1.00	1.00	1.00
InternVL	0.18	1.00	1.00	0.92	1.00	1.00	1.00	1.00

Figure 3: All $8 \times 4 = 32$ model–factor measurements, rounded to two decimals. Darker gray denotes lower correctness; each cell also prints its value. RA ranges from 0 to 1, whereas II ranges from about 0.876 to 1. Both measures condition on a correct audit-base decision.

A.2 All Model–Factor Measurements

Table 3 reports all $8 \times 4 = 32$ model–factor measurements. Each factor has 200 frozen static items and 200 frozen audit triplets. Static estimates use 200 reported outcomes except for Phi-3.5-Vision on Spatial ($n = 198$); missing outputs are not imputed. The conditional denominator n_B counts base-correct audit triplets for each model and factor.

Table 3: Complete model–factor ledger (A_S, RA, and II in percent). Table 5 gives audit-base denominators. Static estimates use $n = 200$ except Phi–Spatial ($n = 198$, marked †).

Judge	Factor	A_S	RA	II
Qwen3-VL-4B	Count	91.50	17.00	100.00
Qwen3-VL-4B	Attribute	82.00	100.00	100.00
Qwen3-VL-4B	Presence	89.50	100.00	100.00
Qwen3-VL-4B	Spatial	80.00	98.50	100.00
Gemma-3-4B	Count	38.50	32.41	87.59
Gemma-3-4B	Attribute	72.50	95.19	97.33
Gemma-3-4B	Presence	75.00	60.10	100.00
Gemma-3-4B	Spatial	56.00	0.00	97.95
Molmo-7B-D	Count	62.50	18.63	97.06
Molmo-7B-D	Attribute	67.00	98.96	98.96
Molmo-7B-D	Presence	79.50	74.37	98.99
Molmo-7B-D	Spatial	66.00	39.86	99.32
Skywork-VL-Reward-7B	Count	76.00	62.81	98.99
Skywork-VL-Reward-7B	Attribute	93.00	100.00	100.00
Skywork-VL-Reward-7B	Presence	84.50	100.00	98.49
Skywork-VL-Reward-7B	Spatial	84.50	86.00	100.00
Idefics3-8B	Count	81.50	71.43	95.77
Idefics3-8B	Attribute	76.50	100.00	100.00
Idefics3-8B	Presence	79.00	100.00	99.50
Idefics3-8B	Spatial	64.00	55.00	100.00
Phi-3.5-Vision	Count	63.50	20.00	95.15
Phi-3.5-Vision	Attribute	80.50	84.95	89.25
Phi-3.5-Vision	Presence	92.00	73.89	97.22
Phi-3.5-Vision	Spatial	$84.85^\dagger$	47.40	98.44
LLaVA-OneVision-7B	Count	78.50	42.08	98.91
LLaVA-OneVision-7B	Attribute	80.50	100.00	100.00
LLaVA-OneVision-7B	Presence	88.50	98.00	100.00
LLaVA-OneVision-7B	Spatial	78.50	87.00	100.00
InternVL3-8B	Count	69.00	18.00	100.00
InternVL3-8B	Attribute	85.50	100.00	100.00
InternVL3-8B	Presence	80.00	100.00	100.00
InternVL3-8B	Spatial	68.50	91.50	100.00

Table 4: Complete natural-image census under $|\Delta A_S| \leq 1$ pp. All gaps are absolute percentage points.

Factor	Pair	ΔA_S	ΔRA	ΔII
Spatial	Skywork–Phi	.35	38.60	1.56
Presence	Molmo–Idefics3	.50	25.63	.51
Presence	Molmo–InternVL	.50	25.63	1.01
Attribute	Phi–LLaVA	.00	15.05	10.75
Presence	Qwen–LLaVA	1.00	2.00	.00
Count	Molmo–Phi	1.00	1.37	1.91
Presence	Idefics3–InternVL	1.00	.00	.50

The original natural-image census has four RA gaps of at least 10 pp, three above 20 pp, and one above 30 pp. Its median RA gap is 15.05 pp and median II gap is 1.01 pp. Phi and LLaVA have

$$\widehat{A}_{S,\mathrm{Attr}}(\mathrm{Phi}) = \widehat{A}_{S,\mathrm{Attr}}(\mathrm{LLaVA}) = 80.5\%, \tag{9}$$

but RA of 84.95% and 100.00%, respectively. The largest original RA gap, Skywork–Phi Spatial (38.60 pp), has a 0.35 pp static difference, but greater relevant-edit pixel change and weaker human/gold agreement limit its interpretation. For Attribute, relevant edits instead have a lower average pixel-change fraction than irrelevant edits (0.118 versus 0.179). These comparisons remain separate from the new same-domain matches.

Table 5: Audit-base accuracy and conditional denominators for the 32 measurements in Table 3. Entries give A_B (%) / n_B.

Judge	Count	Attribute	Presence	Spatial
Qwen3-VL-4B	100.0 / 200	100.0 / 200	99.5 / 199	100.0 / 200
Gemma-3-4B	72.5 / 145	93.5 / 187	99.0 / 198	73.0 / 146
Molmo-7B-D	51.0 / 102	96.5 / 193	99.5 / 199	74.0 / 148
Skywork-VL-Reward-7B	99.5 / 199	100.0 / 200	99.5 / 199	100.0 / 200
Idefics3-8B	94.5 / 189	100.0 / 200	100.0 / 200	100.0 / 200
Phi-3.5-Vision	82.5 / 165	93.0 / 186	90.0 / 180	96.0 / 192
LLaVA-OneVision-7B	91.5 / 183	100.0 / 200	100.0 / 200	100.0 / 200
InternVL3-8B	100.0 / 200	100.0 / 200	100.0 / 200	100.0 / 200

A.3 Full Robustness Tables

Missing outputs. Phi-3.5-Vision reports 168 correct Spatial static outputs among 198 available, or $168/198 = 84.85\%$. Treating its two unavailable outcomes as both errors or both correct gives $168/200 = 84.00\%$ and $170/200 = 85.00\%$; both are within 1 pp of Skywork's 84.50%. The receipts do not identify why the two outcomes are unavailable.

Paired bootstrap. For each primary matched pair, we jointly resample source triplets for both judges (10,000 replicates; seed 20270915). Table 6 reports signed model-A minus model-B contrasts; the sign depends on the printed order. Static contrasts use shared static item identifiers. Their point estimates can therefore differ slightly from differences between separately rounded static percentages.

Table 6: Paired-bootstrap contrasts for all primary matched pairs (percentage points; 95% percentile intervals). Audit contrasts jointly resample 200 matched source triplets for both judges.

Factor	Ordered comparison	Static [95% CI]	RA [95% CI]	II [95% CI]
Spatial	Skywork − Phi	−0.51 [−6.06, 5.05]	38.60 [29.78, 47.23]	1.56 [0.00, 3.61]
Attribute	Phi − LLaVA	0.00 [−5.50, 5.50]	−15.05 [−20.43, −10.21]	−10.75 [−15.46, −6.42]
Presence	Molmo − Idefics3	0.50 [−6.50, 7.50]	−25.63 [−31.66, −19.60]	−0.51 [−2.04, 1.00]
Presence	Molmo − InternVL	−0.50 [−6.00, 5.50]	−25.63 [−31.82, −19.60]	−1.01 [−2.53, 0.00]
Count	Molmo − Phi	−1.00 [−6.50, 4.50]	−1.37 [−10.23, 7.55]	1.91 [−2.68, 6.40]
Presence	Qwen − LLaVA	1.00 [−3.00, 5.00]	2.00 [0.50, 4.00]	0.00 [0.00, 0.00]
Presence	Idefics3 − InternVL	−1.00 [−7.00, 5.00]	0.00 [0.00, 0.00]	−0.50 [−1.50, 0.00]

Shared-base-correct support. Table 7 recomputes RA and II on the intersection of each pair's base-correct sets. This common-support diagnostic does not replace the judge-conditional definitions in Eqs. 2–3. For Molmo–Phi Count, the signed contrast changes from -1.37 pp under judge-specific conditioning to $+13.00$ pp on shared support. The conditioning sets differ, so we retain this as a sensitivity result rather than replacing the primary estimate. Shared support is pair-specific and cannot define a global ranking of judges.

Table 7: All seven primary matched pairs on shared base-correct support. Entries give model-A/model-B percentages; gaps are absolute percentage points.

Factor	Pair	$n_\cap$	RA A/B	$\lvert\Delta\text{RA}\rvert$	II A/B	$\lvert\Delta\text{II}\rvert$
Spatial	Skywork–Phi	192	86.98 / 47.40	39.58	100.00 / 98.44	1.56
Presence	Molmo–Idefics3	199	74.37 / 100.00	25.63	98.99 / 99.50	0.50
Presence	Molmo–InternVL	199	74.37 / 100.00	25.63	98.99 / 100.00	1.01
Attribute	Phi–LLaVA	186	84.95 / 100.00	15.05	89.25 / 100.00	10.75
Count	Molmo–Phi	100	17.00 / 4.00	13.00	99.00 / 100.00	1.00
Presence	Qwen–LLaVA	199	100.00 / 97.99	2.01	100.00 / 100.00	0.00
Presence	Idefics3–InternVL	200	100.00 / 100.00	0.00	99.50 / 100.00	0.50

Table 8: Sensitivity to the within-factor static-matching tolerance. All gaps are absolute percentage points.

Tolerance	Pairs	Median RA	90th RA	Max RA	RA $\geq$ 10 pp
1 pp	7	15.05	30.82	38.60	4
2 pp	11	15.05	25.63	38.60	7
3 pp	15	15.14	34.90	51.64	11

Table 9: Image-level edit diagnostics by factor. Pixel-change fraction and global SSIM use images resized to 256 pixels. SSIM is a global RGB statistic, not a semantic-mask measure.

	Pixel-change fraction		Global SSIM	
Factor	Relevant	Irrelevant	Relevant	Irrelevant
Count	0.229	0.229	0.925	0.922
Attribute	0.118	0.179	0.974	0.958
Presence	0.175	0.201	0.935	0.918
Spatial	0.375	0.265	0.817	0.890

In an additional post-hoc sensitivity analysis, the magnitude-matched subset retains 459 of 800 triplets under $|p_R - p_I| \leq 0.05$: 191 Count, 109 Attribute, 103 Presence, and 56 Spatial. Among the 109 Attribute triplets, Phi is base-correct on 101 and has RA $= 83.17\%$; LLaVA has RA $= 100.00\%$. The restricted-subset RA gap is approximately 16.8 percentage points.

B Completed Validity Controls

This appendix gives the complete independent CLEVR static matched-pair census, the frozen candidate-order diagnostic, and the blinded human perceptual audit.

B.1 Independent Same-Domain Static Census

For transparency, Table 10 reports the complete static panel from which every same-domain match and exact-score class is formed.

Table 10: Complete independent same-domain CLEVR static panel. Each entry is correct / 200 (accuracy in percent). These 32 scores are the only static scores used to form the within-factor 1/2/3 pp same-domain matches and exact-score classes.

Judge	Count	Attribute	Presence	Spatial
Qwen3-VL-4B	200/200 (100.0)	200/200 (100.0)	199/200 (99.5)	200/200 (100.0)
Gemma-3-4B	146/200 (73.0)	193/200 (96.5)	199/200 (99.5)	149/200 (74.5)
Molmo-7B-D	99/200 (49.5)	193/200 (96.5)	193/200 (96.5)	148/200 (74.0)
Skywork-VL-Reward-7B	199/200 (99.5)	200/200 (100.0)	197/200 (98.5)	200/200 (100.0)
Idefics3-8B	186/200 (93.0)	200/200 (100.0)	199/200 (99.5)	200/200 (100.0)
Phi-3.5-Vision	153/200 (76.5)	195/200 (97.5)	185/200 (92.5)	189/200 (94.5)
LLaVA-OneVision-7B	180/200 (90.0)	200/200 (100.0)	197/200 (98.5)	200/200 (100.0)
InternVL3-8B	200/200 (100.0)	200/200 (100.0)	199/200 (99.5)	200/200 (100.0)

On the independent CLEVR static set, the frozen within-factor matching bands yield 42, 44, and 54 pairs at 1, 2, and 3 pp, respectively. Table 11 reports the complete 1 pp census as absolute percentage-point gaps. Static accuracy uses 200 items per model and factor. RA and II retain the original-order intervention judgments and condition on each model's correct audit bases; Table 5 gives their denominators. Pairs sharing a judge are not independent replicates, and pair counts carry no significance claim. Equal measured accuracy at the static ceiling does not establish equal performance on harder unseen items.

Table 11: Complete same-domain CLEVR pair census under $|\Delta A_S| \leq 1$ pp. Short model names match Table 20; gaps are absolute percentage points.

Factor	Pair	ΔA_S	ΔRA	ΔII
Attribute	Gemma–Molmo	0.0	3.78	1.64
Attribute	Gemma–Phi	1.0	10.24	8.08
Attribute	Idefics3–InternVL	0.0	0.00	0.00
Attribute	Idefics3–LLaVA	0.0	0.00	0.00
Attribute	Idefics3–Qwen	0.0	0.00	0.00
Attribute	Idefics3–Skywork	0.0	0.00	0.00
Attribute	InternVL–LLaVA	0.0	0.00	0.00
Attribute	InternVL–Qwen	0.0	0.00	0.00
Attribute	InternVL–Skywork	0.0	0.00	0.00
Attribute	LLaVA–Qwen	0.0	0.00	0.00
Attribute	LLaVA–Skywork	0.0	0.00	0.00
Attribute	Molmo–Phi	1.0	14.02	9.72
Attribute	Qwen–Skywork	0.0	0.00	0.00
Count	InternVL–Qwen	0.0	1.00	0.00
Count	InternVL–Skywork	0.5	44.81	1.01
Count	Qwen–Skywork	0.5	45.81	1.01
Presence	Gemma–Idefics3	0.0	39.90	0.50
Presence	Gemma–InternVL	0.0	39.90	0.00
Presence	Gemma–LLaVA	1.0	37.90	0.00
Presence	Gemma–Qwen	0.0	39.90	0.00
Presence	Gemma–Skywork	1.0	39.90	1.51
Presence	Idefics3–InternVL	0.0	0.00	0.50
Presence	Idefics3–LLaVA	1.0	2.00	0.50

Factor	Pair	ΔA_S	ΔRA	ΔII
Presence	Idefics3–Qwen	0.0	0.00	0.50
Presence	Idefics3–Skywork	1.0	0.00	1.01
Presence	InternVL–LLaVA	1.0	2.00	0.00
Presence	InternVL–Qwen	0.0	0.00	0.00
Presence	InternVL–Skywork	1.0	0.00	1.51
Presence	LLaVA–Qwen	1.0	2.00	0.00
Presence	LLaVA–Skywork	0.0	2.00	1.51
Presence	Qwen–Skywork	1.0	0.00	1.51
Spatial	Gemma–Molmo	0.5	39.86	1.38
Spatial	Idefics3–InternVL	0.0	36.50	0.00
Spatial	Idefics3–LLaVA	0.0	32.00	0.00
Spatial	Idefics3–Qwen	0.0	43.50	0.00
Spatial	Idefics3–Skywork	0.0	31.00	0.00
Spatial	InternVL–LLaVA	0.0	4.50	0.00
Spatial	InternVL–Qwen	0.0	7.00	0.00
Spatial	InternVL–Skywork	0.0	5.50	0.00
Spatial	LLaVA–Qwen	0.0	11.50	0.00
Spatial	LLaVA–Skywork	0.0	1.00	0.00
Spatial	Qwen–Skywork	0.0	12.50	0.00

Only two of the seven original natural-image 1 pp pairs remain within 1 pp on the new CLEVR set: Idefics3–InternVL Presence and Qwen–LLaVA Presence. Phi–LLaVA Attribute differs by 2.5 pp; Molmo–Idefics3 and Molmo–InternVL Presence by 3.0 pp; Skywork–Phi Spatial by 5.5 pp; and Molmo–Phi Count by 27.0 pp. The original seven-pair and new 42-pair medians describe different selected pairs. Their difference does not isolate a domain effect.

B.2 Exact-Score Classes and Count Joint Transitions

Table 12 reports every maximal exact-score class with at least two judges in the same-domain panel. Membership uses equality of the unrounded correct/200 static fractions, not a rounding or tolerance rule. The 12 singleton classes supply no cross-judge comparison and are retained in the machine-readable report. This post-result regrouping re-expresses existing RA; it neither replaces the 42-pair census nor changes its frozen matching bands. The range is the maximum minus minimum judge-conditional RA, not a population or significance claim.

Table 12: All six non-singleton exact same-domain static-score classes. For the two five-judge classes, each member shares all 200 correct bases. Other classes retain judge-specific denominators from Table 5.

Factor	Static (%)	Complete class membership	RA span (pp)
Count	100.0	InternVL, Qwen	1.00
Attribute	96.5	Gemma, Molmo	3.78
Attribute	100.0	Idefics3, InternVL, LLaVA, Qwen, Skywork	0.00
Presence	98.5	LLaVA, Skywork	2.00
Presence	99.5	Gemma, Idefics3, InternVL, Qwen	39.90
Spatial	100.0	Idefics3, InternVL, LLaVA, Qwen, Skywork	43.50

The optional joint Count analysis in Table 13 uses the archived complete binary correctness ledger. Conditional on a correct base, relevant correctness means updating the response and irrelevant correctness means preserving it. The four joint possibilities partition that support; labels describe responses, not internal reasoning. No item, label, gold contract, matching threshold, or primary score is changed.

Table 13: Post-result Count joint transitions: counts / base-correct support. Columns sum to their denominators. Skywork's individual and shared-base supports coincide; one of Qwen's 200 correct bases is outside shared support.

Relevant / irrelevant response	Qwen	Qwen (shared)	Skywork (both)
Update / preserve (selective update)	34/200	34/199	123/199
Retain / preserve (rigid retention)	166/200	165/199	74/199
Update / change (indiscriminate flip)	0/200	0/199	2/199
Retain / change	0/200	0/199	0/199

B.3 Blinded Human Perceptual Audit

The same 320 frozen triplets were labeled by A1 and A2. The two raw submissions and append-only event logs were separately retained and hashed before the blinded relevant/irrelevant mapping was opened. A1 made three corrections, preserved in the event history; A2 made none. No item is missing, excluded, or adjudicated. The second person's distinct identity was user-attested, not independently audited. Table 16 retains all 80 choices per factor and state. A1 used a one-triplet page, whereas A2 used a ten-triplet scroll page; this is a presentation difference, not a different response schema.

Table 14: Full-triplet human audit. All-three-gold requires all three responses to match programmed gold. Exact agreement requires identical A1/A2 labels on all three images, counting unable as a response option. Denominators are 80 per factor and 320 overall.

Factor	A1 all three gold	A2 all three gold	Both all three gold	Exact agreement
Count	65/80	63/80	58/80	61/80
Attribute	67/80	69/80	60/80	62/80
Presence	49/80	52/80	41/80	58/80
Spatial	50/80	37/80	36/80	63/80
All	231/320	221/320	195/320	244/320

Table 15: Post-freeze descriptive transitions relative to each person's own base choice. Relevant flip and irrelevant preserve need not imply a match to programmed gold. Every cell retains 80 triplets per factor.

Factor	A1 flip	A1 preserve	A2 flip	A2 preserve	Both people flip/preserve
Count	72/80	73/80	80/80	71/80	63/80
Attribute	75/80	72/80	77/80	77/80	70/80
Presence	79/80	78/80	74/80	73/80	71/80
Spatial	78/80	78/80	79/80	79/80	76/80
All	304/320	301/320	310/320	300/320	280/320

In Table 15, A1 has 15 relevant retentions, 18 irrelevant changes, and one base-undetermined Count triplet; A2 has 10 retentions and 20 irrelevant changes, with no unable responses. Conditional on a determinate base response, no edited response is unable. A1's one irrelevant-image unable occurs on the same triplet as its base unable and is therefore counted in the base-undetermined transition category. On base-gold-correct support, A1 shows both expected transitions on 231/248 and A2 on 221/237; on base-gold-incorrect support, they do so on 66/71 and 78/83. Thus the transition pattern alone cannot repair base-label disagreement. Both-branch transition-category agreement is 281/320; both people show both expected transitions on 280/320.

Across all 320 triplets, A1/A2 exact agreement is 268/320 for base, 262/320 for relevant, and 260/320 for irrelevant, counting the unable option as a response category. Both give a determinate answer on 319/320, 320/320, and 319/320 states respectively; agreement on that support is 268/319, 262/320, and 260/319. A1 has two unable image responses; A2 has none. At the triplet level, both match all three gold choices on 195/320, and their three response labels are identical on 244/320.

Table 16: Human image-level agreement with programmatic gold. Each cell is gold match / gold mismatch / unable to determine, with denominator 80. Edited-image roles are mapped only after both raw submissions were frozen.

Factor	Annotator	Base	Relevant	Irrelevant
Count	A1	72/7/1	71/9/0	70/9/1
	A2	71/9/0	71/9/0	64/16/0
Attribute	A1	74/6/0	71/9/0	68/12/0
	A2	72/8/0	69/11/0	69/11/0
Presence	A1	51/29/0	50/30/0	49/31/0
	A2	56/24/0	56/24/0	55/25/0
Spatial	A1	51/29/0	51/29/0	51/29/0
	A2	38/42/0	37/43/0	39/41/0

A1 marks all images judgeable on 312/320 triplets (6 no, 2 unsure); A2 on 317/320 (2 no, 1 unsure). They mark obvious artifact, occlusion, or other visual ambiguity on 6/320 and 5/320 triplets. The single combined flag and empty free-text notes cannot identify why human choices diverge from gold. All these human/gold comparisons are descriptive analyses specified after raw-data freezing; no prospective agreement threshold was used.

Scene and response-mapping diagnostic. After observing this disagreement, we replayed the frozen UI loader and checked all 960 displayed-image hashes, all 960 source-scene hashes, candidate contents, and blinded image-role mappings. An independent answer calculation uses only each literal question and the scene's objects and geometry, not its stored answer or preferred label. It reproduces all 960 sampled gold labels. All 240 Spatial answers also agree with the recorded projected horizontal ordering of the queried object centers. These checks detect no scene-label or candidate/image-role mapping discrepancy; they do not independently establish pixel-level visibility or reconstruct the annotators' understanding.

The responses are not uniform across printed answer positions (Table 17). For example, both people match all three Spatial labels on every base-gold-A triplet, but much less often when base gold is B. All sampled Presence and Spatial base answers are semantically *yes*, so this stratification distinguishes A=yes/B=no from A=no/B=yes presentations. Entirely complemented three-state responses account for 29/80 and 21/80 Presence triplets and 28/80 and 41/80 Spatial triplets for A1 and A2, respectively. This is consistent with a candidate-content/printed-label response heuristic, but does not establish one: the same human did not judge each image in both candidate orders, and the quality flags do not record that decision process. We neither invert these labels nor claim to have resolved the human-validity question.

Table 17: Post-result human order diagnostic, retaining all 80 triplets per factor. Cells count all-three-gold triplets within each base-gold-position stratum; A/B here denotes printed position, not answer content.

Factor	A1: base gold A	A1: base gold B	A2: base gold A	A2: base gold B
Count	35/43	30/37	31/43	32/37
Attribute	32/37	35/43	36/37	33/43
Presence	41/44	8/36	39/44	13/36
Spatial	35/35	15/45	35/35	2/45

Joint support is smaller than the individual controls. Table 18 retains the original Count estimates and adds post-result strata based on both humans' three-state gold agreement. The complementary audited stratum is reported, not discarded. The Qwen–Skywork contrast is present in both strata. Human/gold concordance is outcome-dependent selection and does not yield a new representative or prospectively validated benchmark. Only six Count triplets occur in both the human and order samples, five of which are human/gold-concordant. Thus the three controls do not jointly validate all 200 Count triplets. Machine-readable sensitivity results retain every pair at each frozen 1/2/3 pp band, including empty supports as undefined rather than zero. Across all 42 same-domain 1 pp pairs, human/gold-concordant common support leaves 11 absolute RA gaps at least 10 pp and 17

zero gaps (median 2.44 pp). On full-audit common support, the corresponding counts are 16 and 16 (median 2.01 pp). The selected stratum therefore preserves some counterexamples, not a universal or selection-invariant large-gap pattern.

Table 18: Qwen–Skywork Count evidence chain. Fractions are relevant- or irrelevant-correct judgments / base-correct support. Human/gold strata are post-result diagnostics; full-audit estimates are unchanged.

Support and order	Qwen RA	Skywork RA	Qwen II	Skywork II
Full audit, original	34/200	125/199	200/200	197/199
Full audit, shared bases	34/199	125/199	199/199	197/199
Human/gold-concordant, shared bases	12/57	35/57	57/57	56/57
Other audited, shared bases	3/22	12/22	22/22	22/22
Order subset, original	1/20	13/20	20/20	20/20
Order subset, swapped	3/20	13/20	20/20	20/20
Joint human/order support, original	1/5	4/5	5/5	5/5
Joint human/order support, swapped	1/5	4/5	5/5	5/5

B.4 Frozen Candidate-Order Diagnostic

The swap exchanges only the printed positions of the candidate contents; semantic answer content and original-order raw judgments remain unchanged. Each of the eight judges has 240 unique swapped-order judgments, for 1,920 new rows. Table 19 reports all seven original pairs on the frozen 20-triplet-per-factor subset. The descriptive sensitivity labels use no post-result numeric threshold.

Table 19: Original- and swapped-order RA on the frozen 20-triplet subset. Entries give model-A/model-B percentages; each orientation has its own base-correct denominator. "Partial" denotes a persistent direction with a material magnitude change; "Strong" denotes a reversal or loss of comparable support.

Factor	Ordered pair	Original RA A/B	Swapped RA A/B	Sensitivity
Count	Molmo–Phi	36.4 / 16.7	0.0 / 21.4	Strong
Attribute	Phi–LLaVA	89.5 / 100.0	87.5 / 100.0	Stable
Presence	Molmo–Idefics3	75.0 / 100.0	50.0 / 100.0	Partial
Presence	Molmo–InternVL	75.0 / 100.0	50.0 / 100.0	Partial
Presence	Qwen–LLaVA	100.0 / 95.0	100.0 / 100.0	Stable
Presence	Idefics3–InternVL	100.0 / 100.0	100.0 / 100.0	Stable
Spatial	Skywork–Phi	90.0 / 50.0	90.0 / 42.1	Partial

For Phi–LLaVA Attribute, Phi is base-correct on 19/20 and 16/20 triplets in the two orders. Its RA is 87.5% in each order on the 16 triplets common to both base-correct sets; LLaVA is base-correct on all 20, with RA 100% in both. Molmo Count is base-correct on 11/20 and 9/20, but those sets have no intersection; its content-level orientation consistency is 4/60. The Molmo–Phi Count direction is therefore not robust, and a common-support contrast for Molmo is undefined, not zero. In a separate single-model diagnostic, Qwen Count is base-correct on 20/20 in each order, has RA 1/20 versus 3/20 and II 20/20 in both, and has content consistency 58/60. We do not label an average of the two orientations "position-bias-free".

C Implementation Details

C.1 Checkpoints, Prompt, and Decoding

Table 20 lists the eight evaluated checkpoints. Frozen run receipts retain resolved revisions and checkpoint-file hashes. Public repository identifiers make the model identities auditable here without exposing machine-local paths.

Table 20: Evaluated checkpoints and inference adapters. All models serve as binary pairwise judges; Skywork uses a reward-head adapter.

Display name	Checkpoint repository	Adapter
Qwen3-VL-4B	`Qwen/Qwen3-VL-4B-Instruct`	Qwen pairwise
Gemma-3-4B	`google/gemma-3-4b-it`	Transformers pairwise
Molmo-7B-D	`allenai/Molmo-7B-D-0924`	Molmo native pairwise
Skywork-VL-Reward-7B	`Skywork/Skywork-VL-Reward-7B`	Skywork reward head
Idefics3-8B	`AI-ModelScope/Idefics3-8B-Llama3`	frozen expanded pairwise
Phi-3.5-Vision	`LLM-Research/Phi-3.5-vision-instruct`	frozen expanded pairwise
LLaVA-OneVision-7B	`AI-ModelScope/llava-onevision-qwen2-7b-ov-hf`	frozen expanded pairwise
InternVL3-8B	`OpenGVLab/InternVL3-8B-hf`	frozen expanded pairwise

Every non-reward judge received the same frozen prompt. The image was supplied through its model-specific chat-template placeholder:

```
You are scoring a pairwise visual
question-answering preference.
Question: {question}
Candidate A: {candidate_a}
Candidate B: {candidate_b}
Look at the image. Choose the better
answer to the question.
Reply with exactly one token: A or B.
```

The parser accepts a binary A/B decision. Generation is deterministic (`do_sample = False`, temperature unset, at most 16 new tokens). The expanded checkpoint adapter uses bfloat16 and SDPA; Phi uses eager attention, as required by its frozen environment. Skywork scores the two candidates with its value head rather than generating A/B. Candidate order is identical across judges for each item and shuffled deterministically across items.

C.2 Static Preference Set Construction

The original independent static set has 800 natural-image preference problems, 200 per factor. Count draws from TallyQA (Acharya et al., 2019) and pairs a gold count with a nearby incorrect count. Attribute, Presence, and Spatial draw from GQA (Hudson & Manning, 2019); their alternatives use structured attribute substitutions, presence alternatives, and relation inversions or binary alternatives, respectively. Each item stores one gold preference after a deterministic A/B order shuffle.

By construction, these natural images share neither examples nor a selection procedure with the RewardLens-CLEVR audit.

The independent same-domain control follows RewardLens Same-Domain Static Protocol v2.1, not the superseded v2 protocol. It specifies 800 base-only CLEVR items, 200 per factor, with independent semantic and candidate-order schedules and exactly 100 printed gold A and 100 gold B choices per factor. The prospective universe has 40,000 generation attempts (4 factors $\times$ 200 slots $\times$ 50 retry indices), with distinct generation- and render-seed domains. The full-universe collision census

found 40,000 unique seeds and zero collision classes in each domain; any duplicate seed-class members would have been rejected deterministically. Retention required PASS validation, and all 800 items passed at retry index zero. The frozen manifest binds item IDs, question and candidate content, scene JSON, image assets, and canonical scene signatures. All 800 final images and 800 scene files were checked before inference; the dataset freeze receipt preceded model judgments. Available scene-signature and audit-overlap checks passed, but this control cannot certify fields absent from the historical audit.

C.3 REWARDLENS-CLEVR GENERATION AND VALIDATION

The audit contains 200 frozen PASS triplets per factor, generated from CLEVR-style procedural scenes (Johnson et al., 2017). Each triplet has base, relevant, and irrelevant renders: 800 triplets and 2,400 images in total. We retain the first 200 sorted PASS identifiers per factor; borderline and failed triplets are not promoted.

Construction and validation implement the contracts in Sec. 4. Count recounts target-color objects after matched additions. Attribute checks target uniqueness and target or non-target recoloring. Presence recounts the queried color–shape conjunction after matched removals. Spatial recomputes the queried left/right relation from the rendered 3D scene state and checks matched world-space displacement. Every validator also enforces $y_R \neq y_B$ and $y_I = y_B$, together with a shared camera, lighting, and render seed within each triplet. These programmatic checks validate the rendered scene specification.

C.4 ANALYSIS-PLAN PROVENANCE AND DATA RECEIPTS

Analysis-plan provenance. The retained version-controlled analysis plan specifies within-factor accuracy-matching bands of 1, 2, and 3 percentage points. Repository identifiers, timestamps, and cryptographic receipt digests are omitted for double-blind review. We call the document a *frozen analysis plan*, not a formal preregistration.

Data and evaluation receipts. The frozen audit manifest retains 200 validated triplets per factor: 800 triplets and 2,400 rendered images. The independent natural-image static set contains 200 items per factor (800 total), using TallyQA for Count and GQA for Attribute, Presence, and Spatial. Its manifest and integrity receipt remain in the project archive; identifying digests are omitted for double-blind review. Separate v2.1 protocol, implementation, and dataset freezes, a validation report, an inference completion receipt, and eight 800-row judge ledgers document the CLEVR static control. The order control has its own frozen 80-triplet manifest, unchanged original-order ledgers, eight 240-row swapped ledgers, and a completion receipt. Neither control changed the original audit judgments or rematched pairs after observing outcomes. The human audit separately retains a frozen 320-triplet sample, its blinded edit-role assignment, two complete raw submissions and append-only event logs, a raw-data freeze receipt, and a post-freeze descriptive analysis. The annotators' different page layouts and full-denominator disagreements are disclosed in Appendix B.3.

C.5 REPRODUCIBILITY COMMANDS

Exact local paths and command names are omitted for double-blind review. The project archive retains commands for scene generation and validation, manifest freezing, offline inference preflight, metric recomputation from frozen judgments, paired bootstrap, and manuscript compilation. The paired-bootstrap artifact uses 10,000 resamples (seed 20270915), pairing models by source-triplet key. Other archived commands validate the v2.1 dataset and inference completion, reproduce the full same-domain 1/2/3 pp pair censuses, and validate the orientation manifest and swapped-ledger completion. These materials, together with the hash-checked human-audit analysis command, remain available for the archival release package.