COUNTERFACTUAL TWINS AND EXECUTABLE VALIDITY GATES FOR DIAGNOSING SPATIAL REASONING IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark accuracy reports how often a multimodal LLM answers correctly, not whether it answered by looking. We present TWIN-Diag, a spatial-reasoning benchmark built from *counterfactual twins*. Each item pairs a procedurally generated indoor scene with two single-object interventions of identical displacement magnitude, one flipping the ground-truth answer and the other preserving it. Across eleven MLLMs from five families, answers track the image but not the truth. On relative distance (T-5), the one template passing every gate, answers change on 21.6% of controls, where the truth is held, and 24.1% of treatments, where it flips, against 5.0% for a repeated image. Pooled over models, counterfactual sensitivity is at most $+0.069$ (one-sided 95% bound, rooms as the unit of inference), and no model–template cell survives correction. The control's construction matters. In a pre-registered test on real renders of our previous release's pairs, controls sampled to keep the answer changed fewer answers than controls solved to the treatment's exact magnitude in all ten models (room-level $p = 0.019$, Holm-rejected), inflating measured sensitivity. Ours also match on-screen travel within 10%. Nine executable validity gates read only the data and exit non-zero on failure. All were fixed before any model answered the release (78 pairs, 32 rooms), four after models had answered earlier builds. The gates are binding. An extended gate rejected our previous release, and absolute distance (T-2) fails the question-count gate, so it is descriptive. The plan's confirmatory test, an image ablation on six unseen T-2 items, rejects at its design floor ($p = 1/81$), in conflict with that gate. On T-5 the same ablation fails to show that the image improves accuracy. Physical-plausibility defects that no gate checks do not detectably separate the arms. We group the twenty-four defects found in our work into seven failure modes of counterfactual-benchmark construction.

1 INTRODUCTION

A multimodal LLM that scores 67% on a spatial question set has established only that it emits the correct token 67% of the time. Accuracy cannot say whether it did so by inspecting the scene. Many items of widely used vision-language benchmarks are answerable without the image (Chen et al., 2024b), and in indoor scenes object co-occurrence often gives "which is closer to the bed, the bookshelf or the lamp" a defensible answer from priors alone, which is shortcut learning (Geirhos et al., 2020) in benchmark form. The established remedy is to construct items whose answer flips under a controlled change while the confounds stay fixed (Goyal et al., 2017; Kaushik et al., 2020; Thrush et al., 2022). Our *counterfactual twins* do this for spatial questions. In a procedurally generated room, a treatment moves one object so that the true answer flips, and a control moves it equally far without flipping it (Fig. 1).

Applied to eleven MLLMs from five families, the twins give a bounded answer. Answers track the image but not the truth. On relative distance (T-5), the template passing all nine validity gates, answers change on 24.1% of pairs whose truth flips and 21.6% of pairs whose truth is held, against 5.0% for a repeated image. Pooled over models, counterfactual sensitivity is at most $+0.069$ (one-sided 95% bound). No model–template cell survives correction, though a cell's test has 80% power only for differences of 0.20 or more (§5.5).

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: **A counterfactual twin triple.** Two interventions on the same object of a procedurally generated scene S, here the potted plant. The treatment S^+ moves it so that the ground-truth answer changes. The control S^- moves it by a vector of exactly equal magnitude in another direction so that the answer is preserved, with on-screen travel matched within 10%. A model reasoning about the scene should change its answer on S^+ and keep it on S^-. On the release, gate G6 detects no separation between the arms in pixel RMSE, changed-pixel fraction or on-screen displacement (the last by construction). Brightness is adjusted identically across panels.

Counterfactual construction introduces a shortcut of its own. If the edit that flips the answer tends to be larger than the one that does not, "change my answer when the picture changed a lot" scores without any spatial reasoning. Sampling edits and keeping one that preserves the answer (sample-and-filter) leaves this channel open. We instead solve for a control of exactly the treatment's displacement magnitude, also matched on screen within 10%. A pre-registered test on real renders of our previous release's pairs shows that models respond to the difference. Sample-and-filter controls changed fewer answers than exact ones in ten of ten models (§5.3).

Construction does not establish validity, so we write the conditions for interpreting the data as nine gates, programs that read only the data and exit non-zero on failure, and register confirmatory tests before collecting the answers they use (Nosek et al., 2018; Simmons et al., 2011). Reading only data does not mean a gate was fixed in advance. All nine gates were fixed before any model answered the release, but four of them, and G6's on-screen statistic, were written after models had answered earlier builds, whose questions the release largely reuses (§4). The gates are binding. One removed 28% of the pairs we had been reporting on, the extended G6 rejected our previous release, and absolute distance (T-2) fails the question-count gate, so its results are descriptive. We group the twenty-four defects found in our own work into failure modes of counterfactual-benchmark construction (Table 4).

Our contributions are counterfactual twins with exact magnitude matching, screen-matched within 10% (§3), with a pre-registered real-render test of the matching policy (§5.3), and nine executable, dated validity gates, among them pixel-change and screen-motion balance and stable referential binding (§4). They also include a pre-registered protocol reported with every application of its test (Table 2), a bounded measurement of eleven MLLMs with rooms as the unit of inference (§5), and an audit of the builds the gates rejected, the conclusions they overturned and the physical defects no gate checks (§5.7). Figure 2 summarises the pipeline.

2 Related Work

The spatial-reasoning benchmarks VSI-Bench (Yang et al., 2025), VSR (Liu et al., 2023) and BLINK (Fu et al., 2024) find spatial reasoning a weakness of current MLLMs, and SpatialVLM (Chen et al., 2024a) attacks it in training. All draw items from captured scenes, so none can intervene and watch whether the answer follows. Their truth is also only as good as the annotation, and ReVSI (Zhang et al., 2026) finds that questions derived from 3-D scan annotations, as VSI-Bench's are, can carry wrong answers.

Benchmarks inherit deep networks' preference for the easiest cue (Geirhos et al., 2020). VQA v2 (Goyal et al., 2017) paired each question with a complementary image. MMStar (Chen et al., 2024b)

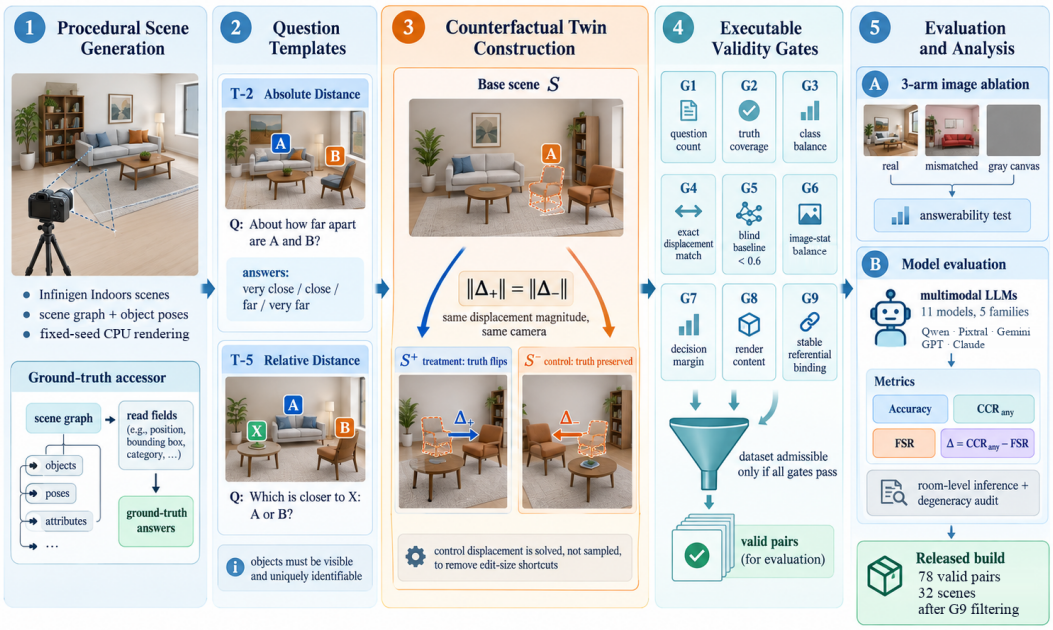


Figure 2: **Overview of TWIN-Diag.** Scenes with exact scene graphs (1), two question templates (2), treatment and control interventions of exactly equal magnitude (3; §3), nine executable validity gates (4; §4) and evaluation with rooms as the unit of inference (5; §5). The room images are illustrations. Figure 1 shows actual renders.

finds visual content unnecessary for many items of existing benchmarks, and SugarCrepe (Hsieh et al., 2023) shows blind models beating sighted ones. Both audit after the fact, whereas we audit at construction time, as an admission gate a failing build cannot pass. This is a procedural answer to Raji et al. (2021)'s argument that construct validity is routinely assumed rather than demonstrated. BetterBench (Reuel et al., 2024) scores benchmarks against a checklist, while our gates are programs run on the data. Miller (2024) argues that questions drawn in groups need clustered standard errors. A room's pairs are such a group, and rooms are our unit of inference.

Counterfactually augmented data (Kaushik et al., 2020), contrast sets (Gardner et al., 2020), Winoground (Thrush et al., 2022) and What's Up (Kamath et al., 2023) rely on human authoring, which bounds scale and leaves the size of the edit uncontrolled. CLEVR-Change (Park et al., 2019) mixes semantic changes with distractor changes of viewpoint and lighting, BEAF (Ye-Bin et al., 2024) edits objects out of images, and EqBen (Wang et al., 2023) tests similarity under minimal image changes. None of them makes the edit that keeps the answer as large as the one that alters it. Two concurrent benchmarks bring the counterfactual to spatial reasoning. CVT-Bench (Vellamcheti et al., 2026) varies the viewpoint, and MindEdit-Bench (Yu et al., 2026), the closest to ours, asks models to predict the consequence of moving an object in photographs, with a human ceiling. Both have the model imagine the edit, whose size is whatever the prompt says. We render both world-lines, so the displacement can be matched and gates can inspect the images. Controlled generators support diagnosis that natural images cannot (Johnson et al., 2017), and SAT (Ray et al., 2025) trains dynamic spatial reasoning, object motion included, in simulated scenes. Infinigen Indoors (Raistrick et al., 2023; 2024) gives the photorealistic procedural scenes and scene graphs we use.

3 Counterfactual Twins

3.1 Scenes and Ground Truth

Scenes come from Infinigen Indoors (Raistrick et al., 2024) via its coarse solver, giving a populated room with an exact scene graph of instances, categories, world-frame poses, axis-aligned extents and support relations, from which ground truth is computed. We render with Blender Cycles at 768×576

on CPU without denoising, since GPU path tracing is not deterministic. On one machine re-rendering reproduces every pixel, so a twin's two images differ only through the intervention. A different CPU changes the noise (66% of pixels differed, E18), so the release was rendered on one host.

3.2 THE INTERVENTION OPERATOR

An intervention relocates one movable object by a displacement Δ and leaves everything else (other objects, materials, camera) in place. Δ is three-dimensional because an object may change its supporting surface, but the answer depends only on horizontal distances. For a question with ground-truth function f we seek a matched pair

$$f(S \oplus \Delta^{+}) \neq f(S), \qquad f(S \oplus \Delta^{-}) = f(S), \qquad \|\Delta^{+}\| = \|\Delta^{-}\|. \tag{1}$$

The third condition matters most, since an approximate match rewards a model for responding to how much the image changed. Sampling controls and keeping the first that preserves the answer (sample-and-filter) yields systematically smaller control edits, and on rendered scenes that changed measured sensitivity (§5.3). We therefore solve for Δ^{-}. When the object keeps its supporting surface, we rotate it at constant height about another object the question names, by the angle $2\arcsin(r/2r_0)$ that makes the chord exactly $r = \|\Delta^{+}\|$, with r_0 its distance to that object. Rotating about the object the answer is measured from leaves the answer unchanged. When the treatment moves it to a new surface, we fix the vertical component to the drop onto the floor and search the horizontal circle that completes the magnitude. The candidate must keep the answer, clear the decision margin and pass the generator's placement checks, and both twins are rendered from the same camera. For the release we also take the candidate whose on-screen displacement is closest to the treatment's and require the two to agree within 10%. If none qualifies, the treatment is re-solved and never approximated (Appendix B.3). Appendix B measures what looser matching would give a rule that reads only $\|\Delta\|$.

The placement checks were weaker than an earlier draft's "collide with nothing and rest on a surface". They tested only other movable objects, never fixed furniture or walls, and accepted anything underneath as support. In a single-rater visual classification by an assistant model (not a person), 27 of the 78 released pairs show interpenetration or implausible support, on one record only in 24 (E25). In 28 pairs the operator also leaves whatever stands on the moved object in mid-air, identically in both twins (E24). Neither defect detectably separates the arms, and no gate checks either (Table 21).

3.3 QUESTION TEMPLATES

We report two templates aligned to VSI-Bench subtasks (Yang et al., 2025). T-2 (absolute distance) asks "About how far apart are the A and the B?" over four bins with edges at 1, 2 and 3.5 m, stated in the prompt. Whether a person can place these distances from one view is untested. Chance is 0.25. T-5 (relative distance) asks "Which is closer to the X: the A or the B?", with chance 0.50. We never compare the two in accuracy. Distance is the horizontal distance between the objects' box centres. Unlike VSI-Bench's prompts, ours do not say "measuring from the closest point of each object". If distances are read as nearest-point gaps between the boxes, 0 of 78 released T-2 states and 10 of 156 T-5 states change label (Table 11, Appendix E.1, E22).

3.4 VISIBILITY AND REFERENTIAL UNIQUENESS

An item is answerable only if each object it names is uniquely identified (Appendix C) and visible above a minimum pixel fraction in a 128×96 ray-cast, the moved object on both intervened world-lines too. A blind answerability probe, with an assistant language model (not a person) answering from the render alone, identified every named object in only 8 of 11 T-5 items. Every failure involved an ordinal phrase. "The rightmost cabinet" ranks the referenceable cabinets, a set a viewer cannot know. We now drop a name group if a same-category object is visible above 30% of the threshold without being a member, and forbid ordinals over more than two objects. Moving one object can still reorder the others, which G9 checks (§4).

4 EXECUTABLE GATES AND PRE-REGISTERED ANALYSIS

4.1 NINE VALIDITY GATES

A template's data are admissible only if they pass all nine checks of Table 1, five released scripts that read only released data, never model output, and exit non-zero on any failure. This does not mean they were fixed in advance. G1–G5 were committed with the analysis plan on 12 August 2026, before any model answered the earlier builds. G6–G9 were written between 4 and 23 September, after models had answered them. G6, G7 and G9 came at an external reader's request, G8 after a model refused to describe a black image (E8), and G9 was re-implemented when it proved to check a proxy (E12). G6's screen statistic was added on 24 September, after an external reader pointed out what G6 missed (E14). A gate written after the answers could have been tuned to them, so for each we report what it changed. For the release all nine were fixed by the release plan before any model answered it (they ran after one model had answered, an order no gate result depends on), but 70 of its 76 questions had been answered on the previous release's renders.

The plan registered that no conclusion is drawn from a template failing a gate in force. Filtering a build that fails a later gate was adopted after the answers for the earlier builds and registered in advance for the release. T-2 fails G1 on the release and is descriptive only. No gate checks physical plausibility (E24, E25). The five templates we specified besides T-2 and T-5 were dropped before any gate existed (Appendix A.1). Of the nine gates (Appendix A), three carry the conceptual weight.

G4 (exact displacement matching) requires every matched pair to satisfy $\|\Delta^{+}\| = \|\Delta^{-}\|$ to within 5×10^{-4} m, which makes counterfactual sensitivity interpretable at all.

G5 (blind baseline < 0.6) requires the best of three text-only strategies (majority class, first option, a category-pair model) to stay below 0.6, a loose threshold for gross leakage. The image-free baseline for the models is the grey-canvas arm of the ablation below.

G6 checks pixel-change and screen-motion balance. G4 matches the world-frame displacement, but a model reads pixels. If the arms were separable in the image, a model responding to how much the picture changed could produce CCR $>$ FSR without tracking the truth. G6 fails a template if pixel RMSE, changed-pixel fraction or the moved object's on-screen displacement separates the arms, by KS $p < 0.05$ or AUC above 0.60 on any one. Combinations are not gated, and a pass means no separation was detected, not that the arms are equivalent.

The other six cover counts, coverage and balance (G1–G3), decision margin (G7), render content (G8) and stable referential binding (G9, every phrase naming the same object on all three world-lines). G7 and G9 re-derive margins and phrases from the released data with the generator's own code.

4.2 THE PRE-REGISTERED ANALYSIS

The analysis plan was committed on 12 August 2026 and amended on 17 and 30 August, after models had answered earlier builds (Appendix D). It fixes one confirmatory test and labels everything else exploratory. Table 2 lists every application of that test. On a build revised after models had answered an earlier version, the release included, the test is applied post hoc (E16). The release plan, a second plan frozen on 24 September before any answer it covers and shipped in anonymised translation, registered the real-render test of §5.3. Its confirmatory test, the ablation on new T-2 items no model had seen, conflicts with G1, which the plan also kept (E21).

We hypothesise that items are answerable from image content. The same model answers the same de-duplicated questions under three arms: the real render, a mismatched render from a pool of one base render per scene, and a constant grey canvas. The pool's fixed indexing showed 16 of its 32 renders (Appendix F). The ordering real $>$ mismatched $>$ constant is fixed a priori, which licenses a one-sided test at $\alpha = 0.05$. With per-item outcome $y_{i,a} \in \{0, 1\}$ and arm weights $w = (2, 1, 0)$, the statistic $T = \sum_i \sum_a w_a \, y_{i,a}$ has an exact null distribution under exchangeability of the arm labels within an item. As an external reader pointed out, it reduces to a paired real vs. grey contrast and does not prove the three-way ordering (Appendix F).

	Criterion	T-2†	T-5
registered with the analysis plan			
G1	distinct questions ≥ 30	25 ✗	51 ✓
G2	truth coverage	3/4 bins ✓	2 values ✓
G3	max. truth share ≤ 0.5	0.480 ✓	n/a
G4	$\|\Delta^+\| = \|\Delta^-\|$	26/26 ✓	52/52 ✓
G5	blind baseline < 0.6	0.480 ✓	0.510 ✓
added after models had answered earlier builds			
G6	arm AUC, pixel RMSE / changed px	0.543 / 0.536 ✓	0.515 / 0.546 ✓
	arm AUC, on-screen displacement	0.501 ✓	0.500 ✓
G7	min. decision margin	0.203 m ✓	0.122 m ✓
G8	renders with content	312/312 ✓	
G9	phrases keep their referents	78/78 pairs ✓	

Table 1: **Validity gates on the release (`merged_r3`).** Every gate covers every record. G6 asks every statistic for KS $p \geq 0.05$ and arm AUC ≤ 0.60 (0.5 is chance). G9 fails on `merged_r3_all` (78/81 pairs); the build reported here passes. G3 is inapplicable to a binary answer space, where G5 carries its protective role. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

4.3 Metrics and the degeneracy audit

A matched pair is a treatment record (truth flips) and a control record (truth held) of equal displacement magnitude. CCR_{any} and *FSR* are the fractions of treatment and of control pairs on which the model's answer changes, over the pairs whose four answers are usable. Sensitivity is $\Delta = \text{CCR}_{\text{any}} - \text{FSR}$. Pairs from one room share its layout and often its camera, so the room is the unit of inference. The test sums the treatment-only minus the control-only answer changes per room and enumerates the exact two-sided sign-flip null. That null, arm exchangeability within each room, allows any dependence inside a room but is stronger than zero mean sensitivity (Appendix H). The 22 model–template tests form one registered family under Benjamini–Hochberg, Holm and Bonferroni, T-2's cells included as descriptive. We adopted the room-level test after an external reader showed our record-level McNemar test could not be rescued by a clustered interval (E11). A cell is flagged as degenerate and excluded from conclusions when one response exceeds 80% of its answers, or when a candidate's excess preference (rate chosen minus rate true) exceeds 0.25 and survives a room-level test at $p < 0.01$, Bonferroni-corrected over the $m = 21$ T-5 candidates (Appendix G).

5 Experiments

5.1 Setup

The release, `merged_r3`, rebuilds the previous release `merged_g9` (101 pairs, Table 23) after G6 was extended to on-screen motion (E14). Every control was solved again to match the treatment's on-screen displacement within 10%: 55 treatments kept their record with a new control, 17 were solved again, and 29 pairs admitted none. Of 9 fresh T-2 pairs on new questions, generated after the release plan was frozen, G9 removed 3. The release has 78 pairs in 32 rooms (52 in 27 for T-5), rendered on one host (E18), with 25 distinct T-2 and 51 T-5 questions. Since 70 of these 76 were asked on `merged_g9`, the release is not an independent replication of it. Eleven models from five families answered both sides of every pair with one prompt and parser on 24–25 September 2026: Qwen2.5-VL-7B (Bai et al., 2025) and Pixtral-12B (Agrawal et al., 2024), both greedy, and three tiers each of Gemini, GPT-5.4 and Claude through their vendors' APIs. One family's temperature cannot be pinned (Appendix J).

5.2 Gate outcomes

T-5 passes all nine gates (Table 1). T-2 fails G1 with 25 distinct questions against 30 (re-solving dropped 15 of its pairs), so no conclusion is drawn from T-2 and every T-2 value below is descriptive. The blind baselines, and the grey arm on T-5, equal the majority-class rate. G6 passes on all three

statistics, but the screen pass is imposed by construction and so is not independent evidence. On `merged_g9` the same statistic failed (AUC 0.727 and 0.632), which is E14 and the reason for the rebuild. G6 gates each statistic alone, yet a responder combining all three, fitted leaving each room out, separates the T-5 arms (AUC 0.587, $p = 0.002$). An in-sample changed-pixel threshold reaches $\Delta = 0.173$ (one-sided $p = 0.047$, uncorrected), as large as the largest T-5 model effect. We argue against this shortcut without ruling it out (Table 6).

5.3 Does exact matching matter on real renders?

The release plan registered one test of the matching policy on renders, D3, on the 101 pairs of `merged_g9` (Table 7). Each control was solved twice more, by sample-and-filter (SF, the generator's first valid non-flipping candidate at any magnitude) and at the worst case of a 0.10 m tolerance (TAU). In one session, ten models (all but Qwen2.5-VL-7B) answered both arms and the re-rendered exact controls (EX). Sample-and-filter controls move answers less, with $D = -58$ over 987 model–pairs in 34 rooms (room-level $p = 0.019$, rejected after Holm), lower for 10 of 10 models. SF would thus understate FSR, and overstate Δ, by about 0.059, the order of the effects in Table 3. TAU is not detected ($p = 0.11$), which does not show a 0.10 m tolerance to be harmless. Construction closes the edit-size channel and filtering does not. The result holds without the pairs with floating objects (E24, $p = 0.018$, rejected after Holm). Without implausible placements (E25) its direction holds and it grows, but at $p = 0.036$–0.045 it is not rejected after Holm (Table 21). EX, the previous release's control, moved further on screen than its treatment, and SF controls are smaller in the world and on screen. D3 therefore tests the policy only on `merged_g9`'s pairs and does not separate world from screen size.

5.4 Is image content used?

The registered test (§4) is answered by Gemini-3-Flash, and Table 2 lists all its applications, recomputed from released answers (E16). It rejected on T-2 on every earlier build. On the release, T-2 fails G1, so 16/8/4 correct of 25 ($p = 1.8 \times 10^{-4}$) is descriptive. The release plan made the six fresh T-2 items, which no model had seen, its confirmatory test. It rejects (4/3/0 correct, $p = 1/81$, the smallest p its design can reach). It also conflicts with the plan's G1 rule (E21), since six items are far below the 30 G1 requires for power. The rejection also rests on the grey arm. T-5, never registered, was tested after it first passed the gates and enlarged after a non-significant run11 ($p = 0.08$), an optional continuation (E17). On the release the test fails to show that T-5 answers use image content (32/21/26 of 51, $p = 0.12$). We withdraw the previous release's value ($p = 0.023$) from the claims (E20). That build's answers give $p = 0.086$ on the questions the release kept. The ablation shows only base renders, which E24 and E25 do not touch (Appendix F).

5.5 Counterfactual sensitivity

Table 3 reports CCR_{any} and FSR per model and template, and Figure 4 shows one model in detail.

Two of twenty-two cells are nominally significant, and none survives correction. Gemini-3.5-Flash reaches $\Delta = +0.173$ on T-5 ($p = 0.048$) and Claude-Opus-5 $+0.269$ on T-2 ($p = 0.039$, descriptive). Benjamini–Hochberg, Holm and Bonferroni over the 22 registered cells retain none, under record-level and equal-room-weight variants as well (Appendix H). Without the pairs affected by E24 or E25, no cell is nominal (Table 21). The test is small. A T-5 cell of 27 rooms has 81% power at $\Delta = 0.20$ and 19% at $\Delta = 0.10$ ($\alpha = 0.05$), and 28% at $\Delta = 0.20$ after Bonferroni, where 80% would need 49 rooms (196 at $\Delta = 0.10$). These powers are upper bounds (Table 20). A non-significant cell has failed to show sensitivity at this size and has not shown its absence.

Pooled over the eleven models, Δ on T-5 is $+0.025$ ($[-0.023, +0.078]$, room bootstrap), so counterfactual sensitivity on T-5 is at most $+0.069$ (one-sided 95% bound, Appendix H.2). The pooled T-2 estimate, $+0.099$ (room $p = 0.028$), is from an unregistered test on a template that fails G1, hence descriptive. Yet answers track the image even where the truth is held. Averaged over models, a T-5 answer changes on 21.6% of controls, 90% of the rate on treatments (24.1%), against 5.0% when the identical base image is asked again (4.6% over both templates). Re-rendering identical geometry with new noise changes 5.9% of Pixtral-12B's greedy answers and 4.0%–17.8% of the API models'

Build	answered	gate status then	T-2 r/m/g (n)	T-2 p	T-5 r/m/g (n)	T-5 p
run5	14 Aug	first build to pass G1–G5 (T-2)	14/5/7 (36)	0.022	–	–
run6	18 Aug	T-2 passed G1–G5; T-5 failed	14/8/3 (30)	0.0012	18/15/14 (32)	0.164*
run7	30 Aug	T-5 only	–	–	19/20/15 (36)	0.205[a]
run11	1 Sep	T-2 failed G1; T-5 passed G1–G5	13/8/6 (23)	0.019*	22/15/16 (33)	0.080
merged (149 pairs)	2 Sep	passed G1–G5	22/9/8 (44)	3.5×10^{-4}	64/45/48 (95)	0.0091
merged_g8 (141)	6 Sep	after G8	22/8/8 (41)	2.6×10^{-4}	61/43/45 (91)	0.0078
merged_g9 (101, previous release)	23 Sep	after G9	17/7/4 (30)	4.4×10^{-5}	43/32/32 (65)	0.023
merged_r3 (78, released)	25 Sep	T-5 passed G1–G9; T-2 failed G1	16/8/4 (25)	1.8×10^{-4}	32/21/26 (51)	0.121
merged_r3 fresh top-up	25 Sep	confirmatory (release plan §5); conflicts with the G1 rule (E21)	4/3/0 (6)	0.012	–	–

Table 2: **Every application of the registered image ablation** (Gemini-3-Flash; correct answers under real / mismatched / grey images (r/m/g) on the de-duplicated questions all three arms parsed, n in parentheses; one-sided exact trend test, registered for T-2 and repeated unregistered on T-5). Every row is recomputed from the released answers with the current parser and de-duplication rule; the dates and gate status are the record's. *Not tested at the time: the template had failed a gate on that build. [a]The record logged real = mismatched = 20/37 (0.541), p = 0.2054, each arm over its own parsed items. The fresh row is the release plan's confirmatory test on the six T-2 items no model had seen. Room-level p and the full record in Table 10.

	accuracy		T-2 sensitivity (26 pairs)†					T-5 sensitivity (52 pairs)				
Model	T-2	T-5	CCR	FSR	Δ	p	audit	CCR	FSR	Δ	p	audit
Qwen2.5-VL-7B	.240	.569	.000	.000	+.000	1.00	**flag**	.212	.154	+.058	.508	pass
Pixtral-12B	.320	.490	.038	.077	−.038	1.00	**flag**	.135	.135	+.000	1.00	pass
Gemini-3.1-Flash-Lite	.600	.549	.692	.500	+.192	.180	pass	.231	.231	+.000	1.00	pass
Gemini-3-Flash	.640	.627	.654	.346	+.308	.063	pass	.327	.327	+.000	1.00	pass
Gemini-3.5-Flash	.640	.608	.615	.500	+.115	.563	pass	.423	.250	**+.173**	**.048**	pass
GPT-5.4-nano	.240	.608	.308	.385	−.077	.688	pass	.212	.154	+.058	.656	**flag**
GPT-5.4-mini	.400	.569	.192	.231	−.038	1.00	pass	.135	.096	+.038	.625	pass
GPT-5.4	.560	.549	.577	.500	+.077	.781	pass	.269	.231	+.038	.781	pass
Claude-Haiku-4.5	.280	.511	.435‡	.304	+.130	.508	pass	.225‡	.200	+.025	1.00	pass
Claude-Sonnet-5	.560	.580	.577	.423	+.154	.289	pass	.216‡	.235	−.020	1.00	pass
Claude-Opus-5	.560	.490	.692	.423	**+.269**	**.039**	pass	.269	.365	−.096	.227	pass

Table 3: **Eleven models, five families** (the release, `merged_r3`). $\Delta = \text{CCR}_{\text{any}} - \text{FSR}$ over pairs with four usable answers; p: two-sided room-level sign-flip test, exact under the arm-swap null (§4; McNemar in Appendix H); audit: the degeneracy verdict per template (Appendix G). Two of 22 cells are nominally significant (bold), none after correction. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule. ‡Usable pairs: Claude-Haiku-4.5 23 (T-2) and 40 (T-5); Claude-Sonnet-5 51 (T-5). 23 answers, all Claude-Haiku-4.5's, say the model cannot see the objects.

(E19, Table 25). Part of what FSR measures is therefore sensitivity to render noise, which should favour neither arm.

5.6 Eleven models, five families

Separate tests do not order models, since a nominal cell beside a null one is not a difference. Gemini-3.5-Flash's Δ exceeds GPT-5.4's on the same T-5 pairs by $+0.135$ ($[-0.06, +0.33]$), and of all 110 same-pair differences 3 are nominal and none survives Holm (Table 18). Nor does tier order the models (Appendix H.1). Three cells are degenerate: both open models on T-2, and GPT-5.4-nano on T-5, which over-selects the potted plant. Across models the potted plant is chosen more often than it is correct (mean excess $+0.163$), but not beyond selection among 21 phrases (max-T $p = 0.068$, against 0.0015 on `merged_g9`, E20). This is exploratory (Appendix G). Only Claude-Haiku-4.5 says it cannot see the objects named (Appendix J).

Failure mode	What went wrong	Entries	n
Leakage and shortcuts	text reveals the answer; controls move further on screen	E1, E14	2
Referential binding	ambiguous phrases; phrases rebind after the move	E2, E12	2
Data and render validity	dark or host-dependent renders; noise drift; truth convention	E8, E18, E19, E22	4
Physical plausibility	floating objects; placements unchecked against furniture	E24, E25	2
Scoring and metrics	incomparable metrics; refusals scored; two de-dup rules	E4, E9, E10	3
Statistical unit and multiplicity	items, builds and rooms treated as independent; BH misstated	E5, E6, E7, E11, E15	5
Reporting and chronology	claims from flawed or enlarged data; chronology; plan conflict	E3, E13, E16, E17, E20, E21	6

Table 4: **The audit as a taxonomy.** The twenty-four registered defects by failure mode (entries in Table 22); eight (E3, E7, E10, E14, E18, E19, E20, E25) are claims we had written that re-measurement did not support.

5.7 What the gates caught

Our gates rejected our own builds and withdrew conclusions we had written down. The registry holds twenty-four defects (E1–E25, with E23 kept in the repository only), grouped in Table 4 by the failure mode each exposes and listed in Table 22 (Appendix I). Fourteen were found by our own re-measurement or review, five by external review, two by a gate, two from outside our process and one by an executable check. Neither a null on a defective build nor a positive on a build enlarged until significant was evidence about T-5 (E3, E17, E20). G6 passed `merged_g9` while its controls moved further on screen in 73 of 101 pairs, which it did not then measure (E14).

6 Discussion

Pooled over eleven models, counterfactual sensitivity on T-5 is at most $+0.069$. On the previous release's pairs, sample-and-filter controls lower FSR in all ten models tested. The confirmatory T-2 test rejects at its design floor, in conflict with G1 (E21). And on T-5 the image moves the answers but has not been shown to make them more accurate (E20).

6.1 Limitations

The release is small, with 32 rooms, 76 questions and 78 pairs, and only T-5 passes all nine gates. After Bonferroni a T-5 cell's test has 28% power at $\Delta = 0.20$, and 80% would take about 49 rooms (Table 20), so a cell without a detected effect is weak evidence of absence. T-5 was never registered and was enlarged after a non-significant run (E17).

Physical plausibility is imperfect. Objects resting on a moved target float in both twins of 28 of 78 pairs (E24), and 27 pairs show interpenetration or implausible support in a single assistant-model rater's classification (E25). Only 31 are defect-free. Neither defect detectably separates the arms (AUC 0.423–0.548, sign tests $p \geq 0.29$), and without the affected pairs no cell is nominal (Table 21). The generator's new checks are off by default, so the release rebuilds byte for byte. A rebuild with them on will be released after review.

The two open models are older than the proprietary ones, no model trained for spatial reasoning is included, and reasoning effort was not controlled. All scenes come from one generator and three room types. Renders reproduce pixel for pixel only on one CPU architecture (E18), and new noise alone moves answers (E19). TWIN-Diag does not rank models.

There is no human baseline yet. The answerability probe used an assistant model, and a blind human study of every released question in all three states is in progress (Appendix C).

To conclude, an evaluation has the right shape only if treatment and control differ in nothing but the property under test, exactly in the world and within 10% on screen. Like a generator's own check, a gate is only as good as its fidelity to the rule it stands for (E12, E25).

Reproducibility statement

We report on the release (`merged_r3`), its version before G9, the previous release (`merged_g9`) and the builds it came from (Table 23), the three control arms of the real-render test (§5.3), and the earlier builds of the execution history (run5, run6, run7 and run11). Every number about these

builds is recomputed from released artifacts by released code. Of the historical figures in the audit trail only E1's come from a build we do not release, and the blind answerability probe of §3 is not released either. The supplementary material holds each build's `ground_truth.jsonl`, `model_view.jsonl` and answer files (thirteen for the release, whose mismatched and grey arms were answered on base images only). It also holds the scene graphs of every scene these builds use and the 128×96 ray-cast segmentations G9 reads, with the earlier builds' segmentations kept apart, and the per-record flags of E24 and E25, including the visual classification. Its code includes the gate scripts and the two that filter a build by G8 and G9, the construction library with the screen-matched control solver, and the analysis scripts, among them `analyze_r3.py` (the release plan's registered tests), `execution_history.py`, `power_bound.py`, `check_carried.py` (E24) and `check_plausibility.py` (E25). Two further scripts draw every figure and generate every appendix table. The package also contains an anonymised English translation of the release plan's pre-registration, with the hashes recorded at each of its freezes. The original plan and its hash will be released after acceptance. The 312 renders of the release (217 distinct images), together with those of its pre-G9 version and of the control arms (516 distinct images in all), and the 377 distinct images of the earlier builds, are at an anonymous link given in the package README, with all metadata stripped and bound to the package by SHA-256 manifests that map each image to its records. A single command (`run_all.sh`) re-runs the nine gates, reporting that T-2 fails G1 on the release as expected, re-runs the release plan's registered tests and the execution history, and regenerates the appendix tables byte-identically with no network access. Another script, `paper_numbers.py`, prints every other number the text cites about these builds. The segmentations and scene graphs were produced from the original scenes by released scripts that need Blender and the Infinigen scenes (about 25 GB), which we do not ship. Scenes come from fixed Infinigen Indoors seeds, but we have not verified that Infinigen regenerates a scene exactly from its seed. Rendering uses Blender Cycles on CPU with denoising disabled, which we verified reproduces every pixel across process invocations and thread counts on one CPU architecture but not across architectures. On a second host 66% of pixels differed (E18). The release was therefore rendered on a single host and answered afresh by every model, and a reader confirms having the pixels the models saw with the SHA-256 manifest, not with a re-render. We do not use GPU path tracing because it is not reproducible even on one machine. The nine gates are five scripts that read only released data, print the table reported as Table 1, and exit non-zero on any failure. They are `check_gates.py` (G1–G5), `check_imgsep.py` (G6, whose on-screen statistic reads the scene graphs), `check_margin.py` (G7), `check_render.py` (G8) and `check_binding.py` (G9). What the pre-registration covers differs by build (Appendix D). Like any build revised after models had answered an earlier version of it, the release has its registered test applied post hoc. The release plan and its construction code were frozen before any image of the release or of the control arms was rendered, and its analysis code before any model answered one. Yet 70 of the release's 76 questions had been answered on earlier renders, and only the six fresh T-2 items had never been seen. Appendix A gives the full gate specifications with thresholds and dates, Appendix E the verbatim prompts and the answer-parsing rule, and Appendix J the per-model denominators, so that every rate in the paper can be checked against a count. One caveat limits reproducibility. The proprietary models are accessed through APIs whose weights may change under a fixed name, so we report query dates (Appendix J).

Ethics Statement

The measurements reported here involve no human participants and no personal data. A blind human study of the released items is in progress (§6). No human data enter this version. Raters take part only with informed consent, and the study's recruitment and ethics review (approval or exemption) will be reported with its results. All scenes are procedurally generated and no captured imagery of real places or people is used, which is one reason the exactly matched intervention this benchmark depends on is possible at all. The proprietary models were queried through their vendors' public APIs under those vendors' terms, and the scene generator and the two open-weight models we run locally are used under their open licences. Compute was modest (CPU rendering and API queries, no model training). Because an API name does not pin a set of weights, we report model identifiers and query dates and do not treat the results as properties of a permanent artifact. The benchmark is diagnostic, and we have tried to keep its claims inside what the design supports. We report the audit trail of our own defects (Table 22) because selective reporting of benchmark construction is the failure mode most likely to mislead a reader here. Two limitations carry ethical as well as methodological weight. First, the benchmark has no human baseline yet (§6). The answerability probe was answered by an

assistant model, not a person, and we describe it that way throughout so that "we could identify" does not imply human performance. Second, a result on this benchmark is evidence about these items under these gates and not about a model's spatial competence in general. A null at this size is a failure to detect at the power we report, not evidence that an effect is absent. We have written the claims to make that scope explicit.

AI USE STATEMENT

Under the authors' direction, generative AI tools were used to help design and implement the experiments and gates, to interpret results, and to translate the pre-registration into English. An assistant model is also the subject of two measurements, the blind answerability probe (§3) and the visual classification of the records flagged for E25 (Appendix H.3), and both are labelled as such. No scene, question or ground-truth label was produced by a generative model. Apart from the two measurements above, every reported number is computed from released data by released code, which `run_all.sh` reruns. The work contains no mathematical claims or proofs. We also used these tools to help with identifying related work, creating figures and polishing the text. The authors reviewed all AI-assisted work and take responsibility for the final content of this work. No LLM is an author.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing Vision-Language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large Vision-Language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 27056–27087, 2024b.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, pp. 148–166, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmov, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

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(11):665–673, 2020. doi: 10.1038/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 (CVPR)*, pp. 6904–6913, 2017.

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 (NeurIPS) Datasets and Benchmarks Track*, pp. 31096–31116, 2023.

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 (CVPR)*, pp. 1988–1997, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9161–9175, 2023.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4623–4632, 2019.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12630–12641, 2023.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen Indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, 2024.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling (COLM)*, 2025.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, pp. 21763–21813, 2024.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632.

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 (CVPR)*, pp. 5238–5248, 2022.

Shanmukha Vellamcheti, Uday Kiran Kothapalli, Disharee Bhowmick, and Sathyanarayanan N. Aakur. CVT-Bench: Probing spatial-state integrity through counterfactual viewpoint transformations. *arXiv preprint arXiv:2603.21114v2*, 2026.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11998–12008, 2023.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. BEAF: Observing BEfore-AFter changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision (ECCV)*, pp. 232–248, 2024.

Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking object-level counterfactual spatial reasoning in VLMs from in-the-wild photos. *arXiv preprint arXiv:2607.00491*, 2026.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3D reasoning. In *International Conference on Machine Learning (ICML)*, 2026.

A GATE SPECIFICATIONS

The nine gates are implemented in five scripts that read the released data and exit non-zero on any failure. Thresholds are given in full with the date each gate was written, and §4 gives the reasoning for the three that carry conceptual weight. G1–G5 were committed with the analysis plan on 12 August 2026 and amended with it on 17 and 30 August (G2's both-ends criterion, G5's per-template baseline and G3's exemption for binary answer spaces). G6 (its two pixel statistics, 4–5 September), G7 (4 September), G8 (6 September) and G9 (21 September, re-implemented 23 September) were written after models had answered earlier builds. G6 gained its third statistic, on-screen motion, on 24 September, after all eleven models had answered `merged_g9` (E14). All nine, with that statistic, were fixed and hashed in the release plan (Appendix D) before any model answered an image of the release. They ran on the release on 25 September, after Qwen2.5-VL-7B alone had answered the 162 records assembled before G9. The gates read no answers, so the order changes no verdict, and Qwen's answers on the pairs G9 removed enter no analysis. Reading only data is not the same as being fixed in advance (E13). The release keeps 70 of its 76 questions and 55 of its treatments from `merged_g9`, which every model had answered.

G1. Distinct questions ≥ 30, counted after de-duplicating by (scene, question text) and not by record. A paired exact McNemar test needs six discordant pairs in one direction to reach $p < 0.05$. At the pilot effect size that puts the minimum at nine items and a defensible floor at thirty. On the release T-5 passes with 51. T-2 has 25 and fails, so by protocol no conclusion is drawn from T-2 on this build, and its results are marked descriptive wherever they appear.

G2. Truth coverage. For templates with a fixed answer space, at least 3 of 4 distance bins must appear among the de-duplicated truths. Templates whose options are generated per item must show both answer positions. On the release T-2 covers 3/4 bins (`very_far` is missing) and T-5 both positions.

G3. Maximum truth share ≤ 0.5, after de-duplication, for fixed answer spaces only (0.480 on T-2). On a binary space with an odd number of questions, as T-5 has here, it is unsatisfiable ($\lceil n/2 \rceil / n > 0.5$), so G5 carries the protection there.

G4. Exact displacement matching. Every matched pair must satisfy $|\|\Delta^+\| - \|\Delta^-\|| < 5 \times 10^{-4}$ m (the tolerance in `check_gates.py`). The tolerance is loose because the construction solves for the magnitude instead of searching for it. The largest deviation in the release is 6.8×10^{-10} m on T-2 and 1.0×10^{-9} m on T-5, the precision at which displacements are stored and not a property of the scenes. At the median object depth of 3.82 m, 5×10^{-4} m is about 0.14 pixels.

G5. Blind baseline < 0.6. Three text-only strategies are fitted on the de-duplicated question set, and the strongest is taken: always answer the majority class, always answer the first option, or a category-pair preference model, fitted leave-one-out, that sees only which object categories the question names. The observed values are 0.480 (T-2) and 0.510 (T-5), each equal to that template's own majority-class rate (the category-pair model reaches 0.353 on T-5). G5 screens the questions and not a model, whose image-free baseline is the grey-canvas arm of the image ablation (Appendix F).

The threshold is loose enough to admit a text-only rule right up to 0.6 on a binary template. It is there to catch gross leakage such as E1, not to certify that text carries no signal.

G6. Pixel-change and screen-motion balance. Three per-record statistics are computed between $I(S)$ and $I(S \oplus \Delta)$. Pixel RMSE and changed-pixel fraction (fraction of pixels differing by more than 18 levels) were both written 4–5 September. The on-screen displacement of the moved object's centre, projected through the render camera on the base and the intervened world-line, was added on 24 September. The gate fails if, on any statistic and per template, a two-sample KS test separates the treatment and control arms at $p < 0.05$ or the AUC of a threshold classifier exceeds 0.60. The pixel statistics were computed a day before their threshold, the $[0.40, 0.60]$ release criterion our leakage audit had used since 4 August, was written into the script. `merged_g9` passed the two pixel statistics but fails the screen statistic, with AUC 0.727 (KS $p = 0.003$) on T-2 and 0.632 ($p = 0.008$) on T-5, because its exact controls travelled further across the image than their treatments (E14). On the release the AUCs are 0.543, 0.536 and 0.501 on T-2 and 0.515, 0.546 and 0.500 on T-5 (RMSE, changed fraction, screen), and no KS test rejects (smallest $p = 0.417$). The screen pass, however, is imposed by the construction, which matches each control's on-screen travel to its treatment's within 10% (Appendix B.3), so it is not independent evidence that the arms look alike. Nor are a KS test that does not reject and an AUC point estimate at or below 0.60 equivalence bounds. Passing means no separation was detected at this n. G6 tests each statistic alone. A rule combining them is not gated and is reported in Table 6. Objects containing an emitter are excluded as intervention targets, because relocating one 152-pixel emitting lamp changed 10.7% of the frame, mostly through cast shadows.

G7. Decision margin (written 4 September). A treatment crossing a bin edge by a centimetre changes the label without changing anything an observer could resolve, and a control stopping a centimetre short invites a noisy observer to switch. Construction requires both world-lines to sit at least 0.20 m (T-2) or 0.12 m (T-5) from the boundary. That constraint lives inside the generator, which a reader cannot check, so the gate re-derives truth and margin from the released scene graphs and refuses to report a margin unless the re-derived labels match the released ones exactly. All 156 of the release's 156 records match. The minimum margins are 0.203 and 0.122 m, so the constraint binds almost exactly.

G8. Render content (written 6 September). Every image must have mean intensity $\geq 20/255$ and at least 20% of pixels above a threshold of 2. The gate exists because the `merged` build failed it, with 32 of 596 renders near-black and one scene at 0.11% non-black pixels (E8). The near-black renders there have mean intensity at most 14.6 and the next darkest 26.1, so any mean threshold between the two removes the same renders. On the release 312 of 312 renders pass (324 of 324 before G9).

G9. Stable referential binding (written 21 September, re-implemented 23 September). Every phrase a question uses must denote the same instance on both intervened world-lines as on the base one. Same-name objects are told apart by their left-to-right order in the frame, so displacing one, or something in front of one, can make "the leftmost chair" name a different chair, or no chair, after the intervention. The construction-time check only asked whether the moved object was still referenceable, which it can be while its phrase has changed. The generator ranks by the centroid of an object's visible pixels in a BVH ray-cast segmentation at 128×96. A released script re-runs the builder's own ray-cast from the original scenes on the base and both intervened world-lines of every record, and the gate re-derives every phrase from those segmentations with the builder's own referring code. Like G7, it first verifies itself, requiring the phrase it derives on the base world-line for every named object to be the one the released question uses. That holds for all 282 records of `merged_g8`, where two runs of the segmentation script agreed record for record, and for all 162 records of the 81-pair rebuild before G9 (0 unverifiable), whose segmentations were made once, on the release's rendering host.

On `merged_g8` (141 pairs), 105 pairs name at least one object by an ordinal, and 40 fail: after the intervention a phrase names a different instance (rebinding) or names none because the group becomes ambiguous or the object drops below visibility. This happens on the treatment world-line only in 13, the control only in 14 and both in 13. Removing these 13 T-2 and 27 T-5 pairs left `merged_g9`. The first implementation ranked projected object centres from the scene graphs, judged only the treatment world-line and only groups whose two members were both named, and had passed all nine groups it could judge on 14 scenes (E12). Of the 81 pairs of the rebuild before G9, 50 are ordinal-exposed and 3 fail, all rebinds on the control world-line only and all among the T-2 top-up pairs. The release keeps 78 pairs, 47 of them ordinal-exposed. G9's judge was also a selection

rule when the rebuild's controls were chosen (Appendix B.3), so the 72 re-solved pairs pass it by construction, and only the 9 top-up pairs met it as a test.

A.1 Which templates reached the gates

Of the seven VSI-Bench-derived templates, only T-2 and T-5 were ever built on real scenes, and they are the only ones a gate has judged. T-2 failed G1 and G2 on an early build (run4c), G1 on run9 and G1 on the release, and T-5 failed G5 on run6 (E1) and G1 on run9. The other five failed no gate, because none reached one. They were set aside while the generators were being written (31 July to 3 August 2026), before the gate script existed (12 August). T-1 (counting) changes its answer only when an object is added or removed, while both samplers propose relocations only. T-3 (object size) needs a swap operator that was never implemented, and T-4 (room size) was declared as a non-flippable control and never built. T-6 (relative direction) was defined from a room-corner observer instead of the rendering camera, so it cannot be answered from the image. For T-7 (route planning), random relocations produced a flip in 0 of 30 attempts, and its answer depends on every object in the room, so no unrelated control can be built. "Two of seven templates pass the gates" would overstate what was tested. Only two of seven were built.

B What exact magnitude matching buys

G4 demands that the treatment and the control of a pair have numerically identical displacement magnitude. That is stronger than the counterfactual literature usually bothers with, and it is fair to ask what it buys. This appendix answers with metadata rules (B.1), with the pre-registered test D3 on rendered scenes (B.2), and describes the release's screen-matched construction (B.3).

B.1 Against metadata rules

Define the *edit-size responder* as a rule that reads nothing but $\|\Delta\|$ and answers "the answer changed" iff it exceeds a threshold. It does no spatial reasoning whatever. Its score is $\Delta(\theta) = P(\|\Delta\| > \theta \mid \text{treatment}) - P(\|\Delta\| > \theta \mid \text{control})$, and the best it can do, maximised over θ, is the one-sided Kolmogorov–Smirnov distance between the two arms' magnitude distributions. Table 5 scores it against control sets that other matching policies would produce. Its first blocks are hypothetical magnitude sets, not generated scenes. Only the bold "as built" rows come from controls that were placed, rendered and validated, and the last block is measured (B.2).

The two arms would differ if nothing forced them not to, and the asymmetry is structural. Flipping a binned-distance answer requires crossing a bin edge, so a treatment displacement has a floor (0.767 m on T-2 in the release), whereas preserving an answer has none, since not moving preserves it. A builder that samples a control and keeps the first valid one therefore produces systematically smaller control displacements. The table's first row asserts the limiting case, and D3 measured the real one. On the T-2 pairs of `merged_g9`, sample-and-filter controls moved 0.503 m on average against 1.074 m for their treatments, and the responder scores 0.800 on them (0.369 on T-5, Table 9). In its worst case, with every control short by exactly τ, a tolerance of 0.10 m still hands the responder 0.192 on T-5, comparable to the largest effects models show (Table 3).

Matching the two arms' magnitude distributions suffices for this responder. As the table shows, any control set whose magnitudes are a permutation of the treatments' scores exactly zero. Per-pair matching adds invariance to rules that combine $\|\Delta\|$ with anything the two twins share (the scene, the view, the question, the moved object). If the twins share $(\|\Delta\|, c)$, any rule $g(\|\Delta\|, c)$ returns the same value on both and scores zero, whereas a permutation guarantees that only for rules of $\|\Delta\|$ alone. The table demonstrates it with $\mathbb{1}[\|\Delta\| > m_0 + \mu]$, where m_0 is the base scene's distance to the decision boundary and μ the construction margin. Against exactly-marginal but shuffled controls it scores $+0.227$ on T-5 and $+0.011$ on T-2, where nearly every magnitude clears the threshold anyway, and it scores zero on both against the per-pair-matched controls. Tightening the match from G4's 5×10^{-4} m to 10^{-9} m buys nothing a model could see. In its worst case G4's tolerance lets the responder separate pairs by half a millimetre (0.14 pixels at the median depth) and score 0.038 on T-2 and 0.019 on T-5. What matters is solving for the control instead of sampling it, which D3 now supports on rendered scenes.

These rules read only metadata. On the release, low-level image change carries little of $\|\Delta\|$ into the pixels (Spearman $+0.073$ with pixel RMSE and $+0.173$ with changed-pixel fraction, $p = 0.031$, over 156 records), but the moved object's displacement across the image tracks it ($\rho = +0.615$). On `merged_g9` that displacement also differed between the arms, which G6 did not then measure (E14). The control moved further on screen in 73 of 101 pairs (mean 261 against 187 pixels at 768×576). On the release it is matched pair by pair (B.3), and the control moves further in 39 of 78 pairs (mean 213 against 214 pixels).

G6 tests each statistic alone. Table 6 fits a responder to all three at once, leaving each room out in turn, and scores it like a model. On `merged_g9` it separated the arms (AUC 0.700 on T-2, 0.636 on T-5) and reached Δ $+0.400$ and $+0.288$, larger than most models' effects. On the release it still separates the T-5 arms somewhat (AUC 0.587, room-swap $p = 0.002$), but its Δ there, $+0.096$ ($p = 0.086$), is below the largest model effect on T-5 ($+0.173$, Gemini-3.5-Flash). On T-2 it scores -0.077. One single-statistic rule is nominal on T-5. Calling the record with the larger changed-pixel fraction the treatment scores 0.173 in sample (swap-null $p = 0.047$ one-sided, 0.093 two-sided) and stays nominal with a held-out threshold. The treatment's fraction is larger in 35 of 52 pairs. None of these is corrected for the rules tried. The scripts are `ablate_matching.py`, `matching_image_space.py` and the robustness module that fills Table 6.

B.2 On rendered scenes: the pre-registered test D3

The release plan (Appendix D) tested the policies of Table 5 on rendered images. For each of the 101 pairs of `merged_g9` the control was solved again under two further policies. *Sample-and-filter* (SF) takes the generator's first valid non-flipping candidate in pool order, with its magnitude left free (100 pairs found one). TAU, the worst case of a 0.10 m tolerance, re-solves the released control's own construction at magnitude $r - 0.10$ m (92 pairs). Candidates were selected by the same probe-time checks as the release (B.3) and rendered on the release's rendering host with the released settings. The ten models other than Qwen2.5-VL-7B answered them in one session (24–25 September 2026), together with the exact controls re-rendered in the same pipeline (EX). A (model, pair) unit counts as changed when the model's answer on the arm's control render differs from its existing answer on the base render. The three arms share that base answer, so its noise does not differ between arms. The primary test compares SF with EX and the secondary TAU with EX, each by a two-sided exact room sign flip on the per-room sums of the model-summed differences, with Holm over the two.

Sample-and-filter controls changed fewer answers than exact ones, with $D = -58$ over 987 (model, pair) units in 34 rooms ($p = 0.0191$, rejected after Holm). FSR was lower under SF for ten of the ten models (Table 8), by 0.059 on average. On the 91 pairs all three arms rendered, pooled FSR is 0.224 under SF against 0.276 under EX. The plan's results log sets this against a projected drop of 0.073. A builder using sample-and-filter would therefore report each model's Δ higher by about 0.059, comparable to most cells of Table 15. No difference was detected for TAU against EX ($D = -23$ over 907, $p = 0.112$, eight of ten models lower). That fails to show a cost of a 0.10 m tolerance at this n but does not show that the cost is zero.

Table 9 gives the plan's descriptives, computed after the answers by a separate module that the two tests do not use (a deviation recorded in Appendix D). SF shortens both the control's displacement in the world and its travel on screen (median ratio to the treatment's 0.557). TAU shortens the world displacement but not the screen travel (1.230). EX, the previous release's control, travels further on screen than its treatment (1.416, E14). SF therefore differs from EX both in the world and on screen, and D3 cannot say which of the two moved the answers. It tests the policy, not a mechanism.

D3 has three further limits. It was run on `merged_g9`'s pairs, whose EX controls are not screen-matched, so it tests the matching policy and not the release's controls. Each arm's Δ in Table 8 takes CCR from the previous release's treatment answers, so only the between-arm FSR contrast comes from one session, and given the answer drift of Table 25 the per-arm Δ values are descriptive. The pixel rows of Table 9 also compare renders from two hosts. The primary result holds on the pairs without floating objects (E24). On the pairs without a visible placement defect (E25) its direction holds and its size grows, but it no longer survives Holm (Appendix H.3).

Table 5: **What exact magnitude matching buys, against metadata rules and on real renders.** Each entry of the first two blocks is the $\text{CCR}_{\text{any}}-\text{FSR}$ of a rule that does no spatial reasoning. The *edit-size responder* answers "changed" iff $\|\Delta\|$ exceeds a threshold; at the best threshold its score is the one-sided Kolmogorov–Smirnov distance between the treatment and the control magnitudes. Only the bold rows are the build (26 T-2 and 52 T-5 pairs). Every other row of these blocks is a hypothetical set of control magnitudes derived from the build's treatment magnitudes r, not a set of generated scenes. The *none* row is asserted, not measured: a control set entirely below every treatment scores 1 by construction. A *worst-case* tolerance τ puts every control at $r-\tau$, the edge of the band that the responder exploits, which no other control set within the band beats (clamped at 0 where $r<\tau$: 5 T-5 pairs at $\tau=0.50$). The *stochastic* row draws each control from $U[r-\tau,\ r+\tau]$ (mean over 5,000 draws; 95th percentile 0.192 on T-2 and 0.096 on T-5); *independently resampled* draws the controls with replacement from the treatment magnitudes (95th percentile of 20,000 draws); *permuted* gives each control another pair's treatment magnitude, so the two marginals are identical although 96.2% of T-2 and 98.1% of T-5 pairs are unequal (maximum over 10,000 permutations). Draws use seed 20260922. Magnitudes are quantised to 1 μm before comparison, 500 times finer than G4's tolerance: the build's within-pair gaps (at most 1.0×10^{-9} m) are storage rounding, and compared unquantised they alone would score 1/26 on T-2 and 1/52 on T-5. The *context responder* also reads m_0, the base configuration's distance to the decision boundary, re-derived from the scene graphs as in G7. A flip must carry the decision variable across the boundary and a further margin (0.20 m on T-2, 0.12 m on T-5), and moving one object by r moves it by at most r, so the rule fires on every treatment (26/26 and 52/52; smallest slack 0.027 m). Per-pair matching gives the control the same $\|\Delta\|$ and m_0, so the rule scores exactly 0. Permuted controls, which the edit-size responder cannot tell from the build, can fall below m_0+ margin only where that threshold exceeds the template's smallest treatment magnitude (3 of 26 T-2 and 39 of 52 T-5 pairs); their row gives the mean and 5th–95th percentile over 5,000 permutations. The random-context row judges the same controls against m_0 shuffled across pairs (mean, 95th percentile in parentheses): what $\|\Delta\|$ alone earns, so the excess over it is what the shared context adds. [a]At G4's own τ the worst case moves every control by 0.5 mm (median 0.14 px on screen; median object depth 3.82 m), and the best threshold separates only the pairs whose treatment magnitudes fall within τ above it: 1 of 26 on T-2 and 1 of 52 on T-5. The entry bounds what G4's check would let through, not leakage in the build. The last block is measured, not derived: the pre-registered test D3 on the 101 pairs of `merged_g9`, not the release (Table 7): ten models answered the controls of two further policies and the re-rendered exact ones (EX) in one session. D/n is the mean change against EX in whether the answer moved from the base render; a negative value lowers FSR and so raises Δ by $|D/n|$. Templates pooled; p: room sign flip.

Control magnitudes	**T-2**	**T-5**
Edit-size responder, $\mathbf{1}[\|\Delta\\| > \theta]$		
none (asserted)	1.000	1.000
worst case, $\tau=0.0005$ m (G4's)[a]	0.038	0.019
worst case, $\tau=0.01$ m	0.115	0.058
worst case, $\tau=0.02$ m	0.192	0.077
worst case, $\tau=0.05$ m	0.269	0.135
worst case, $\tau=0.10$ m	0.385	0.192
worst case, $\tau=0.20$ m	0.654	0.327
worst case, $\tau=0.50$ m	0.962	0.673
stochastic, $\tau=0.10$ m (mean)	0.120	0.074
independently resampled (95th pct.)	0.231	0.154
exact marginal, permuted (max)	0.000	0.000
exact per-pair, as built	**0.000**	**0.000**
max within-pair gap (m)	6.8×10^{-10}	1.0×10^{-9}
Context responder, $\mathbf{1}[\|\Delta\\| > m_0 + \text{margin}]$		
exact marginal, permuted	$+0.011$ [$+0.000$, $+0.038$]	$+0.227$ [$+0.173$, $+0.288$]
random context (m_0 shuffled)	$+0.000$ ($+0.038$)	-0.001 ($+0.077$)
exact per-pair, as built	**0.000**	**0.000**
Measured on real renders (D3; `merged_g9`'s pairs), D/n against EX		
worst case, $\tau=0.10$ m (TAU, 92 pairs)	$-23/907=-0.025$, $p=0.112$	
sample-and-filter (SF, 100 pairs)	$-58/987=-0.059$, $p=0.019$	

Table 6: **A shortcut responder reading only the pair's own renders**, on the previous release (`merged_g9`) and on the release (`merged_r3`). It sees G6's statistics — the moved object's on-screen displacement (log px), the pixel RMSE and the changed-pixel fraction — and answers "changed" for the record it takes to be the treatment; it is scored like a model, $\Delta = \text{CCR}_{\text{any}} - \text{FSR}$ on the matched pairs. The combined rule is a ridge-penalised logistic regression on all three, fitted with each room held out in turn; the single-statistic rules call the record with the larger (or smaller) value the treatment, their threshold fitted the same way. Every p reruns the whole fit with the arm labels swapped a room at a time (2,000 swaps). G6 tests each statistic alone, and a combination of them is not gated.

	`merged_g9`		`merged_r3`	
	T-2	T-5	T-2†	T-5
pairs (rooms)	35 (17)	66 (31)	26 (17)	52 (27)
All three statistics, logistic, fitted leaving one room out				
AUC (room-swap p)	0.700 (0.0070)	0.636 (0.0015)	0.516 (0.324)	0.587 (0.0020)
Δ (room-swap p)	$+0.400$ (5.0×10^{-4})	$+0.288$ (5.0×10^{-4})	-0.077 (0.858)	$+0.096$ (0.086)
One statistic, held-out threshold: Δ (room-swap p)				
log screen displacement px, larger	-0.029 (0.553)	$+0.000$ (0.569)	$+0.000$ (0.992)	$+0.000$ (0.918)
log screen displacement px, smaller	$+0.343$ (0.0095)	$+0.242$ (0.0045)	$+0.000$ (0.934)	$+0.000$ (0.967)
pixel RMSE, larger	-0.057 (0.679)	$+0.030$ (0.418)	$+0.077$ (0.284)	$+0.038$ (0.374)
pixel RMSE, smaller	$+0.057$ (0.380)	$+0.076$ (0.203)	$+0.115$ (0.213)	$+0.019$ (0.463)
changed-pixel fraction, larger	-0.114 (0.930)	-0.061 (0.804)	-0.077 (0.851)	$+0.135$ (0.035)
changed-pixel fraction, smaller	-0.029 (0.576)	$+0.136$ (0.025)	$+0.038$ (0.334)	-0.038 (0.700)

Control arm	pairs	edit size	screen travel	FSR	models below EX	D/n	p vs. EX
exact, as built (EX)	101	.000 / .000	1.42	.276	–	–	–
worst case, $\tau = 0.10$ m (TAU)	92	.406 / .200	1.23	.249	8/10	$-23/907$	0.112
sample-and-filter (SF)	100	.800 / .369	0.56	.224	10/10	$-58/987$	0.019^{h}

Table 7: **Does exact matching matter on real renders?** The pre-registered test D3, on the 101 pairs of `merged_g9`: their controls solved again as sample-and-filter (SF, the generator's first valid non-flipping candidate, magnitude free) and at the worst case of a $\tau = 0.10$ m tolerance (TAU), rendered, and answered by ten models in one session with the exact controls re-rendered (EX). FSR: the answer moved from the base render, on the 91 pairs all three arms rendered, pooled over models. D: over (model, pair) units, answer moves under the arm minus under EX; p: two-sided room sign flip (34 and 33 rooms); ^{h}rejected after Holm over the two. Edit size: the edit-size responder's score (Table 5), T-2 / T-5; screen travel: median ratio of the control's on-screen displacement to its treatment's. No arm is screen-matched, so D3 tests the matching policy, not the release's controls.

B.3 Screen-matched controls

The release keeps G4's exact world magnitude and adds a match on screen (§3). For each pair of `merged_g9` a solver lists every control of the treatment's magnitude that the equal-magnitude constructions produce (the floor circle at 360 angles, or both senses of rotation about each other object the question names) and that passes the generator's checks (answer kept, decision margin, distance convention, placement). It keeps those whose moved object's on-screen travel p_C satisfies $|\log(p_C/p_T)| \le \log 1.10$, with p_T the treatment's, and orders them nearest first. If none qualifies, the treatment itself is re-solved, trying the next flip candidates for the same question and object by increasing magnitude, each with its own screen-matched control. There is no approximate fallback. On the rendering host each list is tried in order, and the first candidate whose moved object stays referenceable (Appendix C) and whose every phrase keeps its referent under G9's judge is rendered. If none passes, the pair is dropped.

The probe rejected 202 control and 165 treatment candidates for visibility, and 25 control and 54 treatment candidates because a phrase rebound (3 further controls lost a referent). 72 of 101 pairs found a candidate. 55 of them keep the released treatment (15 T-2, 40 T-5) with a new control, 17 have both sides re-solved (5 T-2, 12 T-5), and 29 pairs are dropped (15 T-2, 14 T-5). The T-2 top-up (Appendix D) built 9 pairs on questions no earlier build had asked, with the same screen rule, and G9 removed 3, leaving 6. The release has 78 pairs in 32 rooms (3 rooms of `merged_g9` lost

Table 8: **D3 per model and arm** (the 101 pairs of `merged_g9`; Table 7). n: pairs whose base and arm answers both parsed; FSR: the answer on the arm's control render differs from the model's answer on the base render; Δ: $\text{CCR}_{\text{any}}-\text{FSR}$, CCR from the model's existing treatment answers of the same pairs (answered for the previous release, so only the between-arm FSR contrast is from one session) over the pairs whose treatment answers parsed too (Claude-Haiku-4.5: 85, 78 and 86 under SF, TAU and EX); D: the model's answer moves under the arm minus under EX, on the pairs both rendered. Ten of ten models move less under SF than under EX, eight under TAU (one tied).

	sample-and-filter (SF)				**worst case, $\tau = 0.10$ m (TAU)**				**exact (EX)**		
Model	n	FSR	Δ	D	n	FSR	Δ	D	n	FSR	Δ
Pixtral-12B	100	.080	$+0.050$	-8	92	.130	-0.011	-2	101	.158	-0.030
Gemini-3.1-Flash-Lite	100	.240	$+0.100$	-4	92	.250	$+0.098$	0	101	.277	$+0.069$
Gemini-3-Flash	100	.310	$+0.230$	-10	92	.348	$+0.185$	-4	101	.406	$+0.139$
Gemini-3.5-Flash	100	.290	$+0.310$	-5	92	.337	$+0.239$	$+2$	101	.337	$+0.267$
GPT-5.4-nano	100	.210	$+0.050$	-1	92	.152	$+0.098$	-3	101	.218	$+0.040$
GPT-5.4-mini	100	.200	-0.040	-3	92	.185	-0.033	-3	101	.228	-0.069
GPT-5.4	100	.260	$+0.070$	-1	92	.239	$+0.109$	-3	101	.267	$+0.069$
Claude-Haiku-4.5	90	.189	-0.035	-3	80	.175	-0.026	-5	89	.236	-0.081
Claude-Sonnet-5	100	.210	$+0.110$	-19	92	.359	-0.043	-4	101	.396	-0.079
Claude-Opus-5	100	.260	$+0.150$	-4	92	.293	$+0.098$	-1	101	.297	$+0.109$
all models, D/n	$-58/987$, 34 rooms, $p = 0.019$				$-23/907$, 33 rooms, $p = 0.112$				–		
FSR, 91 common pairs, pooled	.224				.249				.276		

Table 9: **What each D3 arm changes** (the plan's descriptives; the 101 pairs of `merged_g9`). Each arm's control is compared with the released treatment of the same pair. Responders score $\text{CCR}_{\text{any}}-\text{FSR}$ of a rule reading only that statistic at its best threshold (one-sided KS distance); AUC and KS p are G6's functions. The world and screen rows are geometry and cover every pair; the pixel rows compare renders from two hosts — the treatments' renders from the previous release against the base and arm renders of the release's rendering host — on the pairs whose view the release re-rendered, so they mix host noise into the contrast.

	SF	**TAU**	**EX**
World: moved-object displacement $\|\Delta\|$ (m)			
pairs, T-2 / T-5	35 / 65	32 / 60	35 / 66
T-2: mean, treatment / control	1.07 / 0.50	1.04 / 0.94	1.07 / 1.07
T-2: controls shorter	30/35	32/32	0/35
T-2: edit-size responder	0.800	0.406	0.000
T-5: mean, treatment / control	0.89 / 0.58	0.91 / 0.81	0.90 / 0.90
T-5: controls shorter	47/65	60/60	0/66
T-5: edit-size responder	0.369	0.200	0.000
Screen: moved object's on-screen displacement (px)			
median ratio, control / treatment	0.56	1.23	1.42
controls moving further	32/100	55/92	73/101
arm AUC (KS p)	0.659 (7.9×10^{-4})	0.596 (0.059)	0.655 (0.0015)
AUC, T-2 / T-5	0.722 / 0.627	0.656 / 0.570	0.727 / 0.632
screen responder	0.280	0.011	0.000
Pixels (pairs whose view the release re-rendered)			
pairs	84/100	77/92	84/101
pixel RMSE: AUC (KS p)	0.592 (0.267)	0.522 (0.669)	0.527 (0.358)
responder	0.155	0.065	0.060
changed-pixel fraction: AUC (KS p)	0.606 (0.095)	0.514 (0.535)	0.516 (0.841)
responder	0.190	0.052	0.036

every pair and 1 was added). All 78 pairs meet the tolerance (largest $|\log(p_C/p_T)| = 0.0947$ against $\log 1.10 = 0.0953$). Every record is a new render on one host. The 55 kept treatments recur with their question, truth and geometry but not their pixels, so `merged_r3` is a re-rendered recomposition of `merged_g9` and not an independent replication of it (E7).

C REFERRING EXPRESSIONS AND STABLE BINDING

Visibility is enforced by ray-casting the scene BVH at 128×96 and requiring each referenced instance to occupy a minimum pixel fraction. Ray-casting replaced an earlier render-based probe that was both slower and prone to non-termination. At construction the intervened object is re-probed on its intervened world-line and the pair discarded if it is no longer referenceable, since the very intervention that flips the answer can push an object out of frame. In the build logs of run13 the check rejected 259 questions after trying up to four candidate pairs each (687 pairs in all), 10.8% of rejections at that stage. On that construction path the other named objects are not re-probed on the intervened world-lines, and G9 covers them. The rebuild's controls were chosen differently, with the probe judging every named phrase on the new world-lines with G9's rule before anything was rendered (Appendix B.3).

The blind answerability probe that found E2 was an assistant language model, not a person. Answering from the render alone, pre-shuffled and blind to ground truth, it recorded whether it could identify every object named. It was given 17 of the 52 blind items from two builds made before the repair and could do so in only 8 of its 11 T-5 items. Every failure involved an ordinal phrase. On a T-2 item it believed it had resolved, it matched "the rightmost cabinet" and "the middle cabinet" to the wrong two of three cabinets, although the two filled 12.9% and 18.7% of the frame, so size was not what fooled it. The sample is small enough that the rate itself carries little weight, but it established the mechanism. Phrases like "the rightmost cabinet" rank the set of referenceable objects, which a viewer cannot know, so with three cabinets in frame the phrase picks out no unique one. The repair drops a name group entirely if any same-category object is visible above 30% of the threshold without being a member, and forbids ordinal expressions over more than two objects. It was never re-tested with the probe.

The repair acts on the base world-line only. Moving an object can reorder a group, as can moving something in front of one, and the construction-time check asked only whether the moved object stayed referenceable. G9 closes that gap with the generator's own rule on both intervened world-lines (Appendix A). On `merged_g8` the 40 pairs it rejected split almost evenly between the arms (13 treatment only, 14 control only, 13 both). On the rebuild before G9 it rejected 3 pairs, all on the control only, and the release has none.

G9 cannot check whether a person reads the frame as the segmentation does, nor whether a person can place the T-2 distances at all. A blind human study of the release is prepared and not yet run, and we report no human numbers. Its unit is the question. Each of the release's 76 questions contributes three states (base, treatment and control), and three Latin-square forms show every question once per form, in a different state on each. No rater therefore sees two states of one question, as no model call does. A rater reads the models' prompt word for word, answers, then clicks each object the question names, with a "can't find / ambiguous" button on offer, so a confident misbinding shows up as a click on the wrong object. The time spent on every item is recorded. Images carry opaque names and no metadata, each rater gets an order of their own, and the answer key is scored by a released script. An earlier packet built on `merged_g8` served as a pilot and is retired.

D THE PRE-REGISTERED PLANS AND THEIR EXECUTION

The first plan was committed on 12 August 2026, while the first build it governs was still being generated and before any model had seen it. Amendments followed on 17 August (per-template truth coverage and blind baseline) and 30 August (G3 inapplicable to binary answer spaces), after models had answered earlier builds but before the merged builds were generated. The plan fixes one confirmatory test, the three-arm ablation on T-2, and labels everything else exploratory. The T-5 ablation uses the identical procedure but was added after T-5 first passed the gates, and is not pre-registered. Every result in §5.5 and §5.6 is exploratory, and so is every choice made after review: the room-level unit of inference (E11), the room-level degeneracy test, and the parser rules R3 and

R4 (E9). Both plans are self-registered. They were committed and hashed in our own repository rather than deposited with a public registry, so their dates rest on our records.

The plan declared in advance that accuracy is reported after de-duplication by (scene, question) and against the majority-class baseline, a flagged template yields no conclusions, and $p \geq 0.05$ is written as "failed to show answerability at $n = N$", never as unanswerability. We adopted this discipline after once switching from McNemar to a trend test because McNemar came back null (p from 0.125 to 0.025). A statistic chosen by its outcome is not evidence, though the same test fixed in advance is.

The registered test was applied to every T-2 build from run5 onward (E16), and Table 10 lists every application, re-scored with the current parser. A result on a build revised after models had answered earlier builds is the registered test applied post hoc, not prospective confirmation. The builds also overlap (E6, E7), so significance on several of them is one result seen several times. T-5 has a further history (E17). It was not tested on run6 ($p = 0.164$, recomputed) and was null on run7 ($p = 0.205$, E3) and on run11 ($p = 0.0801$, $n = 33$). It was then enlarged by merging builds, reaching $p = 0.0091$ on `merged` and 0.023 on `merged_g9`. Continuing to collect after a non-significant result until one is significant inflates the error rate. We disclose it as optional continuation, and the T-5 result rests on the release alone, where it is $p = 0.121$ ($n = 51$), that is, failed to show answerability at $n = 51$ (E20).

Table 10: **The execution history in full.** Each build's three ablation arms, re-scored with the current parser and de-duplication rule; T and the item-level p are the registered trend test, the room-level p permutes the arm labels a whole room at a time. Dates and gate status are the record's. *Not tested at the time (the template had failed a gate on that build); recomputed so the history has no holes. [a]The record logged real = mismatched = 20/37 (0.541), p = 0.2054, each arm over its own parsed items. [b]The RUN6 record logged 6.0e-05 with the parser of the time. The fresh row takes the T-2 records of the release that carry the top-up builder's stamp, the selection analyze_r3 hashed before any answer. The union row pools run5 and run6 as the RUN6 record reported it, with collision-safe pair ids. Rebuilding from the same code on overlapping scenes reproduces questions and renders, so a later build is not an independent replication of an earlier one: run6's T-2 questions share 22 of 47 (scene, question) stems with run5's, 21 on the same base camera (E6); 120 of run11's 120 records recur unchanged in `merged` (E7); 56 of `merged_r3`'s 156 records match a record of `merged_g9` in question, truth and render geometry (55 treatments, 1 control) and 70 of its 76 questions were asked there, 70 on the same base camera.

Build	answered	T	n	r/m/g	T	p (item)	p (room)	rooms	gate status then
run5	14 Aug	T-2	36	14/5/7	33	0.022	0.056	9	first build to pass G1–G5 (T-2)
run6	18 Aug	T-2	30	14/8/3	36	0.0012	9.1×10^{-4}	8	T-2 passed G1–G5; T-5 failed
		T-5	32	18/15/14	51	0.164*	0.188	13	
run7	30 Aug	T-5	36	19/20/15	58	0.205[a]	0.219	11	T-5 only
run11	1 Sep	T-2	23	13/8/6	34	0.019*	0.040	9	T-2 failed G1; T-5 passed G1–G5
		T-5	33	22/15/16	59	0.080	0.061	16	
merged (149 pairs)	2 Sep	T-2	44	22/9/8	53	3.5×10^{-4}	0.0019	22	passed G1–G5
		T-5	95	64/45/48	173	0.0091	0.012	33	
merged_g8 (141)	6 Sep	T-2	41	22/8/8	52	2.6×10^{-4}	0.0016	19	after G8
		T-5	91	61/43/45	165	0.0078	0.011	32	
merged_g9 (101, previous release)	23 Sep	T-2	30	17/7/4	41	4.4×10^{-5}	2.0×10^{-4}	17	after G9
		T-5	65	43/32/32	118	0.023	0.024	31	
merged_r3 (78, released)	25 Sep	T-2	25	16/8/4	40	1.8×10^{-4}	8.3×10^{-4}	17	T-5 passed G1–G9; T-2 failed G1
		T-5	51	32/21/26	85	0.121	0.115	27	
merged_r3 fresh top-up	25 Sep	T-2	6	4/3/0	11	0.012	0.012	6	confirmatory (release plan §5); conflicts with the G1 rule (E21)
run5 + run6 (union)	–	T-2	46	22/8/9	52	7.0×10^{-4}[b]	0.0054	10	

A second plan, the release plan, was written on 24 September, before any new image was probed, rendered or answered. It records the file hashes of three freezes: before the remote probing and rendering, after the render calibration and its contingency, and before the first model answer. An anonymised English translation ships with the supplementary material, and the original and its hash will be released after acceptance. The plan registers D3 (Appendix B.2) with a primary test (SF against EX) and a secondary one (TAU against EX), Holm-corrected, and the screen-matched release (B.3) with all earlier analyses unchanged. It also registers a drift anchor (Table 25), a T-2 top-up

on questions no earlier build had asked, with a stop rule that reads only gate counts, and, as its one confirmatory test of answerability, the registered T-2 ablation run once on the top-up items. Because these items were generated after the test was written and no model had seen them, that is the first prospective application of the registered test (Appendix F).

The plan's log records each deviation and other event with its time. (i) The pre-registered calibration failed because a released control re-rendered on the release's rendering host kept its geometry but not its noise (E18). As the plan's contingency required, every image of the release and every D3 arm, EX included, was rendered on that one host, and every release record was answered again by all eleven models. No answer from an earlier build is reused on the release. (ii) The first launch of the probe script assigned the scene groups to a shell built-in variable, so it did nothing. It was renamed and restarted. (iii) To apply the stop rule, a probe-only pass (the deterministic ray-cast selection without rendering) counted the T-2 questions the re-solve would keep. With the first top-up pass they fell short of 30, so a second pass ran with a new seed. The two passes built 9 pairs, 6 of which pass G9, and together used about 19.9 of the 20 hours of remote time the rule allowed. A third pass takes at least 1.2 hours and could not have closed the gap, so the top-up stopped with T-2 at 25 questions, below G1. The log first recorded 20.6 hours, over the cap, and was corrected against the remote logs. (iv) Long-lived Blender processes that reopened a scene for every frame exhausted the host's memory and slowed rendering by orders of magnitude. They were stopped, and each scene was re-run in a fresh process. Selection is deterministic, so this changed only how the work was executed. (v) A wrap-up script was started twice and ran in duplicate for about three minutes. Both copies were stopped and a single run produced every artefact. (vi) The local incremental assembly first read only the probe logs from before the restart and assembled half of the SF and TAU arms. It was re-assembled from all logs and the missing items answered. The three probe logs agree on every (arm, pair) selection. (vii) Two frozen shell scripts changed after the last freeze. The answering script now points at the G9-filtered release, so that models answer only records that pass G9, and the incremental assembly script got the file pattern that fixed (vi). (viii) The analysis script omitted the plan's per-arm descriptives for D3, which are computed after the answers by a separate module that neither D3 test uses. (ix) The mismatched arm draws the base render of record k from position $2k$ of its pool, so it showed 16 of the 32 pool images, as every earlier build's ablation did. (x) Before the last freeze, two trial calls per vendor, kept in a scratch file and never counted, confirmed that the APIs answered. (xi) Server-overload errors were re-queried under the existing rule that a technical failure, not an answer's content, triggers a re-query.

The plan's §5 names the top-up ablation its confirmatory test, in conflict with its §2 and the first plan's protocol (E21), which give no conclusion for a template that fails a gate. T-2 fails G1 on the release. On the top-up ($n = 6$) the real, mismatched and grey arms get 4, 3 and 0 correct, with $T = 11$ and one-sided $p = 1/81$ (≈ 0.012). This is the largest T these items allow, so the smallest p the design can reach. We report it as the plan's confirmatory test, as §5 says, and state the conflict: six items are far below the 30 that G1's power argument asks for, the result is carried by the grey arm (real and mismatched differ on one item), and it answers only whether the image carries signal the model uses on these six items.

E PROMPTS, ANSWER PARSING AND THE DISTANCE CONVENTION

The prompt is the question text followed by one instruction line, identical across all eleven models and unchanged in the rebuild:

T-2. `Answer with exactly one of: very_close (<1m), close (1-2m), far (2-3.5m), very_far (>3.5m). Reply with only that one word.`

T-5. `Reply with only the name of the object that is closer (exactly as written in the question).`

T-2 answers are parsed by searching for a bin name, and T-5 answers by a *unique identification* rule. The two option phrases are extracted from the question, the content words of each are counted in the reply, and the option with strictly more wins. A tie is unparsed. This replaced a full-phrase containment rule that penalised terse answers. Under it, a model replying `rightmost` to the options "the rightmost chair" and "the leftmost chair" was scored unparsed although its answer was decisive. The replacement was verified to change zero already-parsed answers before it was adopted. A later

refinement, registered before the affected results were inspected, applies the same rule to the first sentence first, because models often answer and then qualify.

Two defects in this parser were found by an independent re-derivation after the earlier builds' results were in (E9), and both are corrected. First, a refusal that names the objects in order to say it cannot find them satisfied the overlap rule and was scored as a substantive answer. Rule R3 treats a reply that states it cannot see, find or identify the objects, or that the image is too dark to, as unusable. Second, the reference object's words leaked into the option match. For "Which is closer to the rightmost cabinet: the leftmost cabinet or the bookshelf?" the reply "The bookshelf is closer to the rightmost cabinet." scores one content word for each option and was recorded unparsed. Rule R4 drops the reference phrase's words from the count, but only when the official rule ties. Both rules read nothing but the reply, the template and the question, and were frozen and hashed before being applied to any answer file.

The form of R4 we declared primary before seeing any effect dropped the reference words unconditionally. It fails this parser's own adoption criterion, that a change must not alter an already-parsed answer. On `merged_g8`, where the reference shares a word with an option, it un-parses 208 decisive answers, 207 of them verbatim option phrases (111 and 110 on `merged_g9`), and would have removed both record-level Benjamini–Hochberg survivors there. We adopted the tie-only form knowing the effect of every variant, and say so. On those builds the two rules changed only Claude-Haiku-4.5. On `merged_g8`, 20 of its answers became unusable (between 19 and 23, depending on how strictly a refusal is recognised) and 1 was recovered. On `merged_g9` they are 18 and 1. Its Δ there moved from -0.152 to -0.156 on T-2 and from -0.067 to -0.074 on T-5, significant under neither. The pre-review parses of those builds ship with them. The release was parsed by the frozen parser from the start (its hash is unchanged in the release plan), so it has no pre-review parse. There R3 marks 23 of Claude-Haiku-4.5's answers unusable (Appendix J).

E.1 The distance convention

T-2's truth is the horizontal (floor-plane) distance between the centres of the two objects' axis-aligned boxes in the scene graph, binned at 1, 2 and 3.5 m. T-5 compares two such distances. Where VSI-Bench's prompt asks for distance "measuring from the closest point of each object", ours drops that phrase and names no convention, so a model is free to read any of them. Two checks keep the labels from depending on that choice. At construction an item is admitted only if a second reading, the gap between the two boxes along the line of centres, gives the same bin or the same order (the generator's convention check). Table 11 also re-reads every state (base, treatment, control) of every released pair under four other definitions. On T-2 neither nearest-point reading changes a label, and the 3-D centre distance changes 1 of 78. On T-5 the 3-D centre distance changes 1 of 156 states and the nearest-point gaps change 10. A model that reads nearer surfaces disagrees with our truth on those states only.

The displacement $\|\Delta\|$ of Appendix B is the 3-D norm, and 119 of the 156 released records move the object vertically by more than 1 mm. The rotation path of the control construction lacked the convention check until the rebuild (E22). In `merged_g9`, 1 of the 202 records fails it (the intervened world-line of one T-5 control), and that pair was re-solved for the release, where 0 of the 156 records fail. On T-2 the tightest state of a pair sits a median 0.179 of its own distance from a bin edge, and 29 of 78 T-2 states have a named object's centre outside the frame, so part of a T-2 distance is inferred rather than seen.

F The image ablation in full

Table 12 gives the per-arm counts on the item set the registered test uses, the registered and room-level trend tests, and the supplementary pairwise contrasts, for all T-2 items of the release, for the six top-up items alone and for T-5 (plotted in Figure 3). T-2 fails G1 on the release, so its all-item column is descriptive, while the top-up column is the release plan's confirmatory test (Appendix D). On the same items, real-image accuracy is above the majority-class baseline on both templates (16/25 against 0.480 on T-2, 32/51 against 0.510 on T-5). On T-2 the grey arm is below the 0.25 chance level (4/25) but the mismatched arm is above it (8/25), so on T-2 an unrelated image is not equivalent to no image. The arms differ on T-5 too, with the mismatched arm below 0.5 (21/51) and the grey arm at

Table 11: **Other readings of "distance"** (the release, `merged_r3`). The truth is the horizontal distance between the two objects' box centres; each state (base, treatment, control) of every pair is re-read under four other definitions, and *kept* counts the states whose released answer (T-2 bin, T-5 choice) is unchanged; *pairs* counts pairs with a changed state; the last columns give the smallest distance of a state's reading to a bin edge or tie (m) and the states closer than the template's margin. The generator's own convention check (centre distance against the gap along the centre line) holds on 156 of 156 records. On T-2 the tightest state of a pair sits a median 0.179 of its own distance from a bin edge (7 of 26 pairs below 0.15, 2 of 26 pairs below 0.1), and 29 of 78 T-2 states have a named object's centre outside the frame.

Reading	kept	pairs	min. margin	below margin
T-2: 78 states of 26 pairs, margin 0.20 m				
horizontal centre distance (the generator's truth)	78/78	0	0.203	0
3-D centre distance	77/78	1	0.024	8
generator centreline gap (binning.surface_dist)	78/78	0	0.002	33
horizontal nearest-point gap of the boxes	78/78	0	0.014	30
3-D nearest-point gap of the boxes	78/78	0	0.014	30
T-5: 156 states of 52 pairs, margin 0.12 m				
horizontal centre distance (the generator's truth)	156/156	0	0.122	0
3-D centre distance	155/156	1	0.078	2
generator centreline gap (binning.surface_dist)	156/156	0	0.004	23
horizontal nearest-point gap of the boxes	146/156	9	0.000	21
3-D nearest-point gap of the boxes	146/156	9	0.000	21

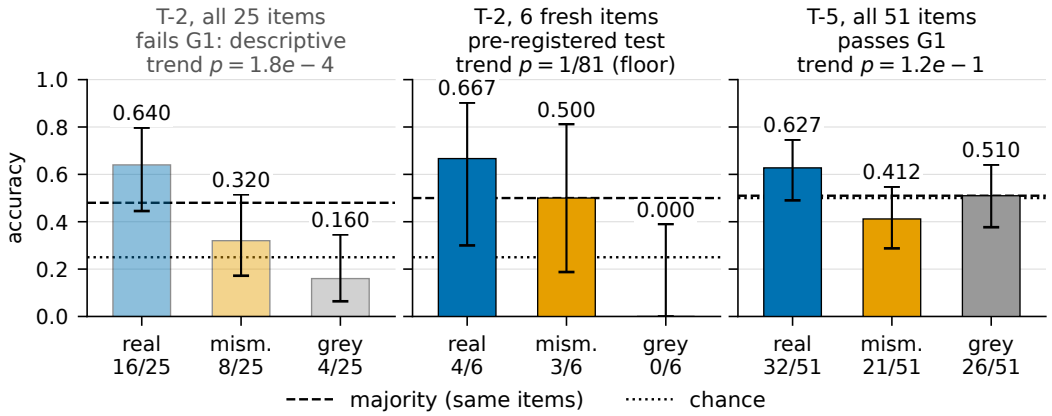


Figure 3: **Three-arm image ablation on the release** (Gemini-3-Flash, the de-duplicated questions all three arms parsed, Wilson 95% intervals). Left: all T-2 items, descriptive because T-2 fails G1. Middle: the six top-up T-2 items no model had seen, the release plan's confirmatory test, at its design floor and in conflict with G1 (E21). Right: T-5, never registered, where the test fails to show answerability. Each p is the registered trend test, a paired real-vs-grey contrast (§4).

the majority rate (26/51). The registered statistic reduces to real against grey (Table 12), and on T-5 that contrast is 14 to 8 ($p = 0.286$). Real against mismatched is 17 to 6 ($p = 0.0347$, unregistered and uncorrected). The T-5 trend test fails to show answerability at $n = 51$ (E20). On the same 51 questions, `merged_g9`'s own answers had given $p = 0.086$, against 0.023 on all 65, so much of the change comes from which questions the rebuild kept.

The mismatched arm's substitutes come from a pool of one base render per scene of the release itself (32 renders, none rejected by G8), assigned by a deterministic index. The arm was asked on the base render only and showed 16 of the pool's images over 156 substitutions. Of these, 6 (3.8%) came from the record's own scene and 2 of those were its own base render, where the arm's answer matched the real arm's in 2/2. Among the items the registered test uses, 1 T-2 item (bedroom_116__t000, a top-up item) and 1 T-5 item received a same-scene substitute and none its own render. Scoring their mismatched answers the other way moves the T-2 trend p from 1.8×10^{-4} to 2.9×10^{-4} and the

Table 12: **The image ablation on one item set.** Gemini-3-Flash answers each item with the real image, a mismatched image or a constant grey image. Items are de-duplicated by (scene, question) over the first record, which is kept only when its base-render answer parsed in all three arms; there are 25 on T-2 and 51 on T-5, and every row uses them. The test is the one-sided exact trend test (registered for T-2; T-5 repeats it unregistered), $T = \sum_{\text{items}}(2\,\text{real} + \text{mismatched})$ with the three arm labels permuted within each item and the null enumerated by convolution; the *per room* row permutes them a whole room at a time (n rooms). The pairwise contrasts are exact McNemar tests, two-sided and supplementary: they were not registered and are not corrected for multiplicity. b counts items the first-named arm answers correctly and the second does not, c the reverse. The trend statistic reduces to a real-vs-grey contrast: per item $2\,\text{real} + \text{mismatched} = (\text{real} - \text{grey}) + (\text{real} + \text{mismatched} + \text{grey})$, and no permutation changes the second term, so the observed T is fixed by real vs. grey's $b - c$ (12 on T-2, 4 on the fresh items, 6 on T-5), and the mismatched arm enters its null only through how many items are informative, i.e. have one or two of the three arms correct (17 of 25 on T-2, 4 of 6 fresh, 33 of 51 on T-5). The *fresh* column is the release plan's one confirmatory test: the six T-2 items generated in the rebuild and never shown to a model before, one in each of six rooms. With these items' outcomes no permutation gives a larger T than the observed 11, so $p = 1/81$ is the smallest this item set can reach. The plan makes it the confirmatory test, and the same plan's G1 rule gives T-2 on this build no conclusion; the two rules conflict (E21). The mismatched arm (asked on the base render only) replaced each image by one of the 32 base renders of the pool it drew from, 16 of which it showed over 156 substitutions; 6 came from the record's own scene (2 its own render, where the arm's answer matched the real arm's in 2/2), and none showed a render G8 rejects. On the items above, one item of T-2 and one item of T-5 had such a substitute (one of these is a fresh item); scoring their mismatched answers the other way gives $p = 2.9 \times 10^{-4}$ on T-2, $p = 0.025$ on the fresh items and $p = 0.121$ on T-5. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

	T-2†			**T-2, fresh**			**T-5**		
Correct answers on the items									
real	16/25			4/6			32/51		
mismatched	8/25			3/6			21/51		
grey	4/25			0/6			26/51		
Trend test, one-sided	T	n	p	T	n	p	T	n	p
per item	40	25	1.8×10^{-4}	11	6	0.012	85	51	0.121
per room (sensitivity)	40	17	8.3×10^{-4}	11	6	0.012	85	27	0.115
Pairwise McNemar, two-sided	b	c	p	b	c	p	b	c	p
real vs. grey	12	0	4.9×10^{-4}	4	0	0.125	14	8	0.286
real vs. mismatched	10	2	0.039	1	0	1.000	17	6	0.035
mismatched vs. grey	7	3	0.344	3	0	0.250	8	13	0.383

top-up p from 0.012 to 0.025, and leaves T-5 at 0.12. The real-against-mismatched contrasts move from 0.039 to 0.092 (T-2) and from 0.035 to 0.023 (T-5). On `merged_g9` the pool had been drawn from the unfiltered build and served near-black renders. An earlier draft asserted that no contaminated item fell in the tested sets, and a later one that contamination could not reach the registered p because the statistic ignores the mismatched arm. The first was wrong. The second does not follow, since the arm enters the null through the count of informative items (Table 12), so a contaminated item can move p either way, as the top-up item does here. Neither floating objects nor placement defects (Appendix H.3) reach the ablation, because its three arms show only the base render, which no intervention touches.

G Degeneracy audit in full

A cell is screened when some candidate's excess preference exceeds 0.25 over candidates offered at least eight times, or when a single response exceeds 80% of the model's answers. A screened excess flags the cell only if it survives a room-level sign-flip test at $p < 0.01$, Bonferroni-corrected over the $m = 21$ distinct T-5 candidate phrases of the release. The test flips the signs of the per-room sums of picked-minus-true jointly and is exact if each room's residual is symmetric. Instances come from

records whose two answers both parse, so a matched pair contributes four answers covering three images, the base render answered twice.

Three of the 22 cells are flagged. On T-2 both open models collapse onto `close`, and on T-5 GPT-5.4-nano is flagged for the potted plant (Table 13). Gemini-3.1-Flash-Lite's potted-plant excess, which flagged on `merged_g9`, is not flagged on the release. Treating each answer as an independent binomial draw, as an earlier draft did, flags 7 of 22 cells. Counting each distinct image and question once and testing at the room level flags the same three as the room rule. With k rooms the smallest one-sided sign-flip p is 2^{-k}, and $21 \cdot 2^{-12} < 0.01 < 21 \cdot 2^{-11}$, so after the correction a phrase can flag only if it is offered in at least 12 rooms, all leaning the same way. Only the potted plant is offered that widely (12 rooms), and GPT-5.4-nano's excess is positive in all of them.

Table 13: **Every screened degeneracy excess.** A cell is screened when one response exceeds 80% of its answers (T-2 rows) or a candidate's excess preference exceeds 0.25 over at least eight offers (T-5 rows). *picked* is how often the model chose the candidate where it was offered, *truth* how often it was correct there, over instances from records whose two answers both parse; *rooms* gives the number of rooms whose sum of picked-minus-true is positive, negative and zero. The flag is decided by the one-sided room-level sign-flip test, exact if each room's residual is symmetric, Bonferroni-corrected over the $m = 21$ T-5 candidate phrases, at $p < 0.01$ (*room* column; $*$ = flags). For comparison, *unique* applies the same test counting each distinct image and question once, and *binom* is the binomial on the instances an earlier draft used, which treats them as independent. The single-response share counts every parsed answer. At the room level three cells are flagged (Qwen2.5-VL-7B T-2, Pixtral-12B T-2 and GPT-5.4-nano T-5); counting unique images, three; by the binomial on the instances, seven. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

Model	T	Screened	picked	truth	rooms	room	unique	binom
Qwen2.5-VL-7B	T-2†	`close` (single)	104/104			share $1.000 > 0.80$		$*$
	T-5	desk lamp	8/8	0.500	2/0/0	1.000	1.000	0.082
		potted plant	54/68	0.515	10/1/1	0.072	0.082	3.7×10^{-5} $*$
		chair	12/12	0.583	1/0/0	1.000	1.000	0.033
Pixtral-12B	T-2†	`close` (single)	91/104			share $0.875 > 0.80$		$*$
	T-5	desk lamp	8/8	0.500	2/0/0	1.000	1.000	0.082
		chair	12/12	0.583	1/0/0	1.000	1.000	0.033
Gemini-3.1-Flash-Lite	T-5	potted plant	57/68	0.515	10/0/2	0.021	0.021	5.1×10^{-7} $*$
		cubby shelf	7/8	0.500	2/0/0	1.000	1.000	0.738
		chair	11/12	0.583	1/0/0	1.000	1.000	0.312
Gemini-3-Flash	T-5	large shelving unit	10/12	0.417	2/0/1	1.000	1.000	0.085
		desk lamp	8/8	0.500	2/0/0	1.000	1.000	0.082
GPT-5.4-nano	T-5	large shelving unit	11/12	0.417	3/0/0	1.000	1.000	0.010
		potted plant	64/68	0.515	12/0/0	$0.0051*$	$0.0051*$	3.5×10^{-13} $*$
GPT-5.4-mini	T-5	floor lamp	17/24	0.417	2/1/0	1.000	1.000	0.078
		chair	12/12	0.583	1/0/0	1.000	1.000	0.033
GPT-5.4	T-5	large shelving unit	12/12	0.417	3/0/0	1.000	1.000	5.8×10^{-4} $*$
		chair	11/12	0.583	1/0/0	1.000	1.000	0.312
Claude-Haiku-4.5	T-5	potted plant	46/56	0.518	10/1/1	0.338	0.267	4.7×10^{-5} $*$
		chair	12/12	0.583	1/0/0	1.000	1.000	0.033
Claude-Sonnet-5	T-5	cubby shelf	8/8	0.500	2/0/0	1.000	1.000	0.082
Claude-Opus-5	T-5	chair	11/12	0.583	1/0/0	1.000	1.000	0.312

Table 14 gives the two phrases that drew the most attention. We withdraw the earlier draft's sign test over eleven models. The models answered the same instances, so their signs are not independent draws. With rooms resampled jointly for all models, the across-model mean excess for the potted plant is $+0.163$ $[+0.044, +0.299]$ on the release, raw one-sided $p = 0.014$, and $p = 0.068$ after the max-T adjustment for having picked it out of 21 phrases. On `merged_g9` it was $+0.190$ with adjusted $p = 0.0015$. On the release it does not survive the selection (E20), and we treat it as exploratory. Within a family the excess falls from the smallest model to the largest in Gemini and GPT but not in Claude (Table 18).

H ROOM-LEVEL INFERENCE

The paired McNemar test treats matched pairs as independent. Pairs from one room are not, since they share its layout, its objects and often its camera. Table 15 gives both tests for every cell, on the

Table 14: **Over-selected T-5 phrases, per model and across models.** Counts are the degeneracy audit's instances, taken only from the records whose two answers both parse: both the base and the intervened answer of such a record count, so a base render is counted once per record that answers it. *chosen* is picked/offered, *truth* the rate at which the phrase was in fact correct on the same instances, and *excess* their difference. Intervals resample rooms (10,000 resamples at seed 20260920); in the mean row one resample of rooms is applied to every model at once, because the models answered the same images and are not independent. An interval is shown only for a phrase offered in at least five rooms (– otherwise): over fewer rooms the resamples take a handful of distinct values and their percentiles are not an interval, and the sign-flip test, exact under per-room sign symmetry, is the evidence — with k rooms the smallest attainable one-sided sign-flip p is 2^{-k} ($2^{-2} = 0.250$ for the two rooms of *the desk lamp*, which its raw p attains). Tests flip the sign of picked − truth per room, jointly for all models. The raw p is exact under per-room sign symmetry; the adjusted p refers the phrase's studentized across-model statistic to the maximum over all 21 candidate phrases of the build (max-T, 400,000 Monte Carlo draws at seed 20260920 for the part beyond the raw p), because these phrases were singled out for looking over-selected. *The potted plant* is offered in 17 questions across twelve scenes; its across-model mean excess is $+0.163$ $[+0.044, +0.299]$, raw one-sided $p = 0.014$ and selection-adjusted $p = 0.068$. *The desk lamp* is offered in two questions across two scenes; its across-model mean excess is $+0.239$, raw one-sided $p = 0.250$ and selection-adjusted $p = 0.898$. Per model, four of the eleven *potted plant* intervals lie above zero.

	"the potted plant"				**"the desk lamp"**			
Model	chosen	truth	excess	95% CI	chosen	truth	excess	95% CI
Qwen2.5-VL-7B	54/68	0.515	+0.279	[+0.14, +0.42]	8/8	0.500	+0.500	–
Pixtral-12B	43/68	0.515	+0.118	[−0.13, +0.35]	8/8	0.500	+0.500	–
Gemini-3.1-Flash-Lite	57/68	0.515	+0.324	[+0.19, +0.46]	6/8	0.500	+0.250	–
Gemini-3-Flash	32/68	0.515	−0.044	[−0.22, +0.17]	8/8	0.500	+0.500	–
Gemini-3.5-Flash	31/68	0.515	−0.059	[−0.20, +0.10]	6/8	0.500	+0.250	–
GPT-5.4-nano	64/68	0.515	+0.426	[+0.35, +0.52]	2/8	0.500	−0.250	–
GPT-5.4-mini	51/68	0.515	+0.235	[−0.04, +0.46]	6/8	0.500	+0.250	–
GPT-5.4	41/68	0.515	+0.088	[−0.17, +0.33]	6/8	0.500	+0.250	–
Claude-Haiku-4.5	46/56	0.518	+0.304	[+0.09, +0.48]	3/6	0.500	+0.000	–
Claude-Sonnet-5	38/68	0.515	+0.044	[−0.14, +0.28]	6/8	0.500	+0.250	–
Claude-Opus-5	40/68	0.515	+0.074	[−0.13, +0.27]	5/8	0.500	+0.125	–
mean of 11 models			+0.163	[+0.04, +0.30]			+0.239	–

same pairs, with a room-bootstrap interval. An earlier draft combined the record-level McNemar test, Benjamini–Hochberg correction, and an uncorrected room-bootstrap interval as a further filter. That composite has no guaranteed error rate, and the filter never removed anything, so in practice it was Benjamini–Hochberg on p-values that assume independence (E11).

For the primary test, let $d_i = \text{changed}_{\text{treat},i} - \text{changed}_{\text{ctrl},i}$ for pair i and $D_g = \sum_{i \in g} d_i$ for room g. The sign flip enumerates all 2^k sign patterns of $(D_1, \ldots, D_k)$. It is exact under the *arm-swap* null: within every room, exchanging the treatment and control labels of all its pairs at once leaves the joint distribution of the model's answers unchanged, so each D_g is symmetric about zero. The pairs of a room may depend on one another in any way. That null implies zero mean sensitivity but is stronger, since a model whose response to the two arms differs in shape but not in mean violates it. It is a behavioural assumption, plausible only when the arms are matched in what a model can see, and D3 shows that models' answers do respond to the control's edit (Appendix B.2). World and screen matching make it plausible but do not guarantee it. Table 16 therefore also tests every cell under the mean-zero null with cluster-robust alternatives, and Table 17 simulates their size.

No cell on the release survives Benjamini–Hochberg, Holm or Bonferroni at any level of analysis. At the record level one cell is nominal (Gemini-3-Flash T-2, $p = 0.021$). At the room level two are (Gemini-3.5-Flash T-5; Claude-Opus-5 T-2, with $p = 0.048$ and 0.039), and weighting rooms equally gives the same two. The room-level p is smaller than the record-level one in some cells and larger in others (Table 15). Three intervals exclude zero (Table 15), and those on T-2 are descriptive. On

`merged_g8` we could not detect within-room over-dispersion of the discordances, and there the room-level test was the more conservative one (4 nominal cells under McNemar, 2 with rooms as the unit), mainly because flipping whole rooms costs power. We adopt it because independence across a room's pairs cannot be assumed, even though dependence was not shown. Every alternative in Table 16 finds 3 nominal cells, and none survives any correction. In the six simulated processes of Table 17 the sign flip rejects no more often than its nominal level, while CR1 and CR3 over-reject in several, most under a skewed room effect. This is evidence about these processes and does not prove size under every mean-zero null. With S rooms contributing, the smallest attainable sign-flip p is 2^{1-S}, so many cells cannot reach Bonferroni's threshold whatever the data (Table 16).

Table 15: **Counterfactual sensitivity per cell, with room-level inference.** Every quantity is computed on the matched pairs whose two records both have a usable answer on both sides (n). b counts pairs in which only the treatment record's answer changed, c those in which only the control's did, and $\Delta = (b - c)/n$. p_{rec} is the exact McNemar test, which treats pairs as independent. p_{room} is the primary test: a two-sided sign-flip test on the per-room sums of $d = \text{changed}_{\text{treat}} - \text{changed}_{\text{ctrl}}$, flipping the sign of every pair in a room together and enumerating all 2^k patterns; it is exact under the arm-swap null (§4) and, its null distribution being discrete, conservative. The interval resamples rooms with replacement (10,000 resamples at seed 20260920, 2.5th and 97.5th percentiles); k is the number of rooms contributing pairs to the cell. Three of the 22 intervals exclude zero ($*$: Gemini-3-Flash T-2, Gemini-3.5-Flash T-5 and Claude-Opus-5 T-2). The bounds printed as 0.000 for Qwen2.5-VL-7B T-2 and Gemini-3.1-Flash-Lite T-2 are exactly zero, an attained value of the discrete resampling distribution; such an interval does not exclude zero. Over the 22-cell family, at the record level one cell is nominally significant ($p < 0.05$); none survives Benjamini–Hochberg at FDR 0.05, Holm or Bonferroni at 0.05; at the room level two cells are nominally significant ($p < 0.05$); none survives Benjamini–Hochberg at FDR 0.05, Holm or Bonferroni at 0.05. Benjamini–Hochberg is a step-up procedure: it rejects the k smallest p for the largest k with $p_{(k)} \le k \cdot 0.05/22$. At the room level no ordered $p_{(k)}$ meets its threshold (closest: $p_{(3)} = 0.063$ against 0.0068); at the record level no ordered $p_{(k)}$ meets its threshold (closest: $p_{(1)} = 0.021$ against 0.0023). Weighting every room equally instead (sign flips on per-room mean d, a sensitivity analysis), two cells are nominally significant ($p < 0.05$); none survives Benjamini–Hochberg at FDR 0.05, Holm or Bonferroni at 0.05. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule; its cells stay in the registered 22-cell correction family.

Model	**T**	n	b	c	Δ	p_{rec}	p_{room}	**95% CI (rooms)**	k
Qwen2.5-VL-7B	T-2†	26	0	0	+0.000	1.000	1.000	[+0.000, +0.000]	17
	T-5	52	7	4	+0.058	0.549	0.508	[−0.056, +0.172]	27
Pixtral-12B	T-2†	26	1	2	−0.038	1.000	1.000	[−0.167, +0.100]	17
	T-5	52	4	4	+0.000	1.000	1.000	[−0.096, +0.116]	27
Gemini-3.1-Flash-Lite	T-2†	26	7	2	+0.192	0.180	0.180	[+0.000, +0.421]	17
	T-5	52	8	8	+0.000	1.000	1.000	[−0.135, +0.157]	27
Gemini-3-Flash	T-2†	26	9	1	+0.308	0.021	0.063	[+0.080, +0.500] $*$	17
	T-5	52	8	8	+0.000	1.000	1.000	[−0.140, +0.145]	27
Gemini-3.5-Flash	T-2†	26	8	5	+0.115	0.581	0.563	[−0.115, +0.400]	17
	T-5	52	14	5	+0.173	0.064	0.048	[+0.035, +0.317] $*$	27
GPT-5.4-nano	T-2†	26	2	4	−0.077	0.688	0.688	[−0.259, +0.107]	17
	T-5	52	8	5	+0.058	0.581	0.656	[−0.111, +0.212]	27
GPT-5.4-mini	T-2†	26	3	4	−0.038	1.000	1.000	[−0.194, +0.143]	17
	T-5	52	3	1	+0.038	0.625	0.625	[−0.038, +0.111]	27
GPT-5.4	T-2†	26	6	4	+0.077	0.754	0.781	[−0.160, +0.385]	17
	T-5	52	7	5	+0.038	0.774	0.781	[−0.089, +0.178]	27
Claude-Haiku-4.5	T-2†	23	6	3	+0.130	0.508	0.508	[−0.125, +0.421]	15
	T-5	40	5	4	+0.025	1.000	1.000	[−0.111, +0.143]	23
Claude-Sonnet-5	T-2†	26	6	2	+0.154	0.289	0.289	[−0.043, +0.400]	17
	T-5	51	6	7	−0.020	1.000	1.000	[−0.130, +0.102]	27
Claude-Opus-5	T-2†	26	9	2	+0.269	0.065	0.039	[+0.074, +0.500] $*$	17
	T-5	52	4	9	−0.096	0.267	0.227	[−0.226, +0.022]	27

Table 16: **The 22 cells under cluster-robust alternatives** (the release, `merged_r3`; two-sided). The registered test is the room sign flip, exact under the arm-swap null (§4); the alternatives test a mean-zero null instead: CR1 and CR3 with $t(G-1)$, CR2 with Bell–McCaffrey degrees of freedom, the wild cluster bootstrap-t with Rademacher (R) and Webb (W) weights (2,000,000 draws for Webb; Rademacher enumerated where 2^G allows), and the room bootstrap, percentile and bootstrap-t (10,000 resamples), all at seed 20260925. S: rooms with a nonzero sum, so the smallest attainable sign-flip p is $p_{\min} = 2^{1-S}$. Bold: below Bonferroni's 0.0023. The bottom rows count the family's survivors under each test, and the degeneracy flags each gives when it replaces the room sign flip of the degeneracy audit. On T-2, 11 of 11 cells cannot reach Bonferroni's threshold whatever the data, 3 cannot reach 0.05. On T-5, 6 of 11 cells cannot reach Bonferroni's threshold whatever the data, 1 cannot reach 0.05. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

Model	T	Δ	rooms	S	$p_{\min}$	sign flip	CR1	CR2	CR3	wild R	wild W	boot %	boot-t
Qwen2.5-VL-7B	T-2†	+0.000	17	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	T-5	+0.058	27	9	.0039	.508	.327	.332	.328	.367	.328	.388	.330
Pixtral-12B	T-2†	−0.038	17	3	.250	1.00	.574	.578	.576	.500	.583	.783	.606
	T-5	+0.000	27	8	.0078	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gemini-3.1-Flash-Lite	T-2†	+0.192	17	9	.0039	.180	.098	.105	.096	.145	.097	.111	.084
	T-5	+0.000	27	11	9.8×10^{-4}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gemini-3-Flash	T-2†	+0.308	17	8	.0078	.063	.013	.019	.017	.047	.024	.019	.024
	T-5	+0.000	27	11	9.8×10^{-4}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gemini-3.5-Flash	T-2†	+0.115	17	8	.0078	.563	.404	.418	.422	.547	.416	.452	.411
	T-5	+0.173	27	14	1.2×10^{-4}	.048	.025	.029	.025	.034	.028	.025	.022
GPT-5.4-nano	T-2†	−0.077	17	6	.031	.688	.423	.432	.431	.531	.433	.519	.425
	T-5	+0.058	27	10	.0020	.656	.495	.501	.501	.461	.503	.578	.512
GPT-5.4-mini	T-2†	−0.038	17	5	.063	1.00	.660	.666	.668	.688	.671	.810	.674
	T-5	+0.038	27	4	.125	.625	.315	.325	.326	.500	.358	.440	.337
GPT-5.4	T-2†	+0.077	17	9	.0039	.781	.601	.616	.625	.777	.612	.646	.611
	T-5	+0.038	27	9	.0039	.781	.576	.581	.582	.691	.586	.686	.579
Claude-Haiku-4.5	T-2†	+0.130	15	9	.0039	.508	.346	.351	.342	.398	.324	.386	.328
	T-5	+0.025	23	7	.016	1.00	.710	.714	.714	.500	.704	.848	.713
Claude-Sonnet-5	T-2†	+0.154	17	8	.0078	.289	.191	.200	.194	.281	.178	.193	.185
	T-5	−0.020	27	6	.031	1.00	.741	.746	.748	.656	.760	.881	.754
Claude-Opus-5	T-2†	+0.269	17	9	.0039	.039	.031	.040	.037	.039	.024	.0082	.046
	T-5	−0.096	27	11	9.8×10^{-4}	.227	.140	.146	.140	.173	.136	.155	.129
					nominal ($p < 0.05$)	2	3	3	3	3	3	3	3
					BH	0	0	0	0	0	0	0	0
					Holm	0	0	0	0	0	0	0	0
					Bonferroni	0	0	0	0	0	0	0	0
					degeneracy flags (of 22)	3	4	3	3	3	4	9	3

H.1 Model differences and the previous release

Since comparing two models' Δ from separate tests does not order them, Table 18 tests the difference on the pairs both models can use, with the same room sign flip. Across all 110 model pairs and templates, 3 differences are nominal and none survives Benjamini–Hochberg or Holm. The largest on T-5, Gemini-3.5-Flash against Claude-Opus-5 (+0.269, $p = 0.023$), is one of them. We make no claim that any model is more counterfactually sensitive than another. Nor does tier order the models within a family. Accuracy rises strictly with tier in 1 of 6 family–template combinations, and Δ in 2.

The release plan asked for every cell on `merged_g9` and on the release side by side (Table 19). The two builds share questions but not pairs, so a row is not a paired comparison and neither column replicates the other. 4 cells were nominal on `merged_g9` and 2 are on the release, 1 on both. The sign of Δ differs in 11 of 22 cells. Gemini-3.5-Flash on T-5 moved from +0.318 ($p = 0.0026$), measured against controls that travelled further on screen, to +0.173.

H.2 Power, and a bound on the pooled effect

A cell that is not significant has failed to show sensitivity at its n. Table 20 simulates, on the release's own rooms (not registered), how large an effect the primary test could have shown. It lets every cell use all of its template's pairs and treats the pairs of a room as independent. Both choices favour the

Table 17: **Size of the room tests under mean-zero nulls** (the release, `merged_r3`): rejection rates over 20,000 simulated data sets per process, on the build's own rooms (T-2: 17 rooms, 26 pairs; T-5: 27 rooms, 52 pairs). These nulls are not the arm-swap null on which the sign flip is exact: *exchangeable p=0.3*: $p_t = p_c = 0.3$ in every room (H_0^{swap} holds); *mixed 0.8/0.2*: fixed T- and C-leaning rooms, p 0.8/0.2 and balancing; *mixed 0.6/0.4*: fixed T- and C-leaning rooms, p 0.6/0.4 and balancing; *skew q=0.9, rooms≥2*: skewed room effect in rooms of ≥ 2 pairs, q = 0.9; *skew q=0.6, rooms≥2*: skewed room effect in rooms of ≥ 2 pairs, q = 0.6; *skew q=0.9, all rooms*: skewed room effect in every room, q = 0.9. flip_{eq} weights rooms equally. The two nominal levels are 0.05 and Bonferroni's over the family; Monte Carlo standard errors 0.0015 and 0.0003.

	T-2					T-5				
null	flip	flip_{eq}	CR1	CR2	CR3	flip	flip_{eq}	CR1	CR2	CR3
nominal $\alpha = 0.05$										
exchangeable p=0.3	0.0224	0.0357	0.0606	0.0529	0.0575	0.0294	0.0464	0.0563	0.0505	0.0534
mixed 0.8/0.2	0.0003	0.0022	0.0006	0.0004	0.0004	0.0001	0.0018	0.0001	0.0001	0.0001
mixed 0.6/0.4	0.0143	0.0294	0.0404	0.0333	0.0367	0.0177	0.0333	0.0333	0.0291	0.0314
skew q=0.9, rooms≥2	0.0232	0.0346	0.0621	0.0520	0.0569	0.0355	0.0461	0.0647	0.0581	0.0608
skew q=0.6, rooms≥2	0.0323	0.0345	0.0919	0.0772	0.0810	0.0432	0.0467	0.0816	0.0732	0.0757
skew q=0.9, all rooms	0.0249	0.0365	0.0591	0.0500	0.0541	0.0338	0.0462	0.0607	0.0540	0.0563
nominal $\alpha = 0.002273$										
exchangeable p=0.3	0.0004	0.0006	0.0038	0.0024	0.0040	0.0010	0.0019	0.0031	0.0020	0.0028
mixed 0.8/0.2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
mixed 0.6/0.4	0.0001	0.0003	0.0011	0.0006	0.0011	0.0004	0.0010	0.0015	0.0009	0.0013
skew q=0.9, rooms≥2	0.0005	0.0008	0.0034	0.0020	0.0032	0.0015	0.0022	0.0048	0.0036	0.0043
skew q=0.6, rooms≥2	0.0007	0.0007	0.0132	0.0080	0.0112	0.0018	0.0021	0.0124	0.0097	0.0114
skew q=0.9, all rooms	0.0008	0.0014	0.0047	0.0029	0.0044	0.0011	0.0019	0.0043	0.0029	0.0040

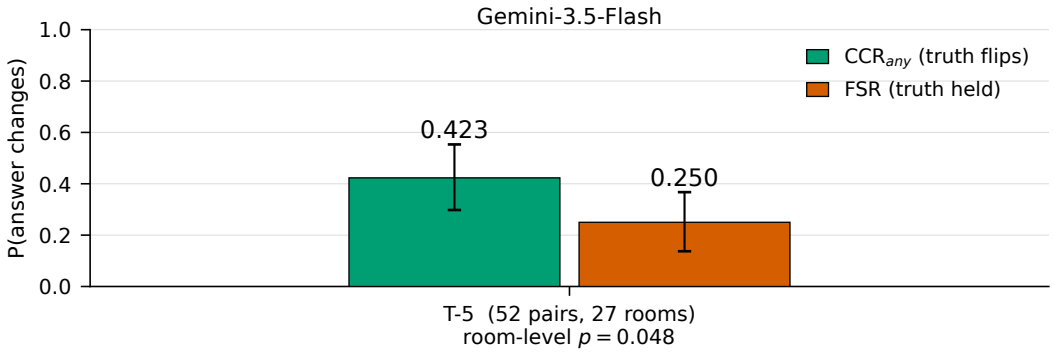


Figure 4: **Counterfactual sensitivity against answer instability** for Gemini-3.5-Flash on T-5, the one T-5 cell that is nominal at the room level ($p = 0.048$). We show it because its effect is the largest on T-5, not because it is typical. It survives no multiplicity correction (Table 15), and its same-pair differences from other models pass no correction either (Table 18). Matched pairs have identical displacement magnitude and on-screen travel matched within 10%, so edit size cannot produce the gap. G6 detects no separation of the arms on any of its three statistics, and a combination of them is not gated (Table 6). Intervals resample rooms, and p is the room-level sign flip, recomputed from the same files that draw the bars.

test, so its powers are upper bounds. In the ρ rows a room's pairs correlate, and power falls. The observed within-room correlation is -0.08 on T-5 and $+0.14$ on T-2. The values 0.2 and 0.5 are illustrations, not estimates. On T-5 (27 rooms, 52 pairs) the test has 81% power to detect $\Delta = 0.20$ at $\alpha = 0.05$ but 28% after Bonferroni over the 22 cells, and 19% and 2% at $\Delta = 0.10$. Eighty per cent power needs $\Delta \geq 0.20$ uncorrected and 0.28 after correction, the latter only by raising the discordance to Δ (Table 20). After correction it would take 49 rooms (about 94 pairs, 1.8 times the release) at $\Delta = 0.20$ and 196 rooms at $\Delta = 0.10$. T-2 is descriptive and weaker still, needing $\Delta \geq 0.33$ uncorrected, with the discordance raised. A non-significant cell is therefore uninformative about effects below about 0.20. The release can find large effects in single models but not small ones.

Table 18: **Model differences on the same pairs** (the release, `merged_r3`; unregistered and exploratory). Each contrast is $\Delta_1 - \Delta_2$ on the matched pairs usable for both models: the room sign flip on per-room sums (p_{room}), its equal-room-weight variant (p_{eq}) and a room-bootstrap interval (10,000 resamples at seed 20260920); $>/<$ counts the pairs where model 1's pair difference is larger/smaller. Benjamini–Hochberg and Holm over these eight contrasts reject none. Tier trends: within a family, the tier labels permuted per room (two-sided, exact); top−bottom is the mean over the pairs all tiers can use of the top tier's pair difference minus the bottom's; *adjacent* is each tier against the one below it. The phrase rows apply the same test to the excess of a salient phrase, per offered instance. Across all 110 model pairs × templates, three are nominal (2 on T-2, 1 on T-5) and none survives BH or Holm. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

Contrast	n	rooms	Δ_1	Δ_2	$\Delta_1 - \Delta_2$	$>/<$	p_{room}	p_{eq}	95% CI
T-5									
Gemini-3.5-Flash − GPT-5.4	52	27	+0.173	+0.038	+0.135	14/6	.253	.130	[−0.06, +0.33]
Gemini-3.5-Flash − Claude-Opus-5	52	27	+0.173	−0.096	+0.269	19/8	.023	.013	[+0.07, +0.48]
Claude-Opus-5 − GPT-5.4	52	27	−0.096	+0.038	−0.135	6/11	.235	.178	[−0.33, +0.05]
Gemini-3.5-Flash − GPT-5.4-mini	52	27	+0.173	+0.038	+0.135	13/6	.119	.090	[+0.00, +0.29]
T-2† (fails G1: no conclusion)									
Gemini-3.5-Flash − GPT-5.4	26	17	+0.115	+0.077	+0.038	7/7	1.00	.883	[−0.29, +0.36]
Gemini-3.5-Flash − Claude-Opus-5	26	17	+0.115	+0.269	−0.154	4/6	.438	.219	[−0.42, +0.13]
Claude-Opus-5 − GPT-5.4	26	17	+0.269	+0.077	+0.192	8/3	.273	.434	[−0.07, +0.40]
Gemini-3.5-Flash − GPT-5.4-mini	26	17	+0.115	−0.038	+0.154	9/7	.465	.453	[−0.14, +0.48]

Tier trend	**T**	n	rooms	top−bottom	p	adjacent: difference (p_{room})	
Gemini	T-5	52	27	+0.173	0.052	+0.000 (1.000)	+0.173 (0.063)
GPT	T-5	52	27	−0.019	0.900	−0.019 (1.000)	+0.000 (1.000)
Claude	T-5	40	23	−0.050	0.661	−0.025 (1.000)	−0.078 (0.432)
pooled	T-5	144	77	+0.042	0.419	–	–
Gemini	T-2†	26	17	−0.077	0.738	+0.115 (0.563)	−0.192 (0.500)
GPT	T-2†	26	17	+0.154	0.329	+0.038 (1.000)	+0.115 (0.613)
Claude	T-2†	23	15	+0.174	0.379	+0.043 (1.000)	+0.115 (0.375)
pooled	T-2†	75	49	+0.080	0.422	–	–

Phrase tier trend	**phrase**	instances	rooms	top−bottom	p
Gemini	the potted plant	68	12	−0.382	5.1×10^{-4}
GPT	the potted plant	68	12	−0.338	0.0060
Claude	the potted plant	56	12	−0.232	0.085
pooled	the potted plant	192	36	−0.323	2.2×10^{-6}
Gemini	the desk lamp	8	2	+0.000	1.000
GPT	the desk lamp	8	2	+0.500	0.222
Claude	the desk lamp	6	2	+0.000	1.000
pooled	the desk lamp	22	6	+0.182	0.272

Averaging Δ over the eleven models helps, with one resample of rooms applied to every model because they answered the same images. On T-5 the average is $+0.025$ $[-0.023, +0.078]$ (room $p = 0.392$), with a one-sided 95% upper bound of $+0.069$ ($+0.070$ without Qwen2.5-VL-7B). Pooled over these eleven models, counterfactual sensitivity on T-5 is at most $+0.069$ at 95% confidence, while the largest single-model bound is $+0.295$ (Gemini-3.5-Flash). On T-2 the pooled Δ is $+0.099$ with an interval above zero and room $p = 0.0284$. This pooled test was not registered and T-2 fails G1, so it is descriptive and we draw no conclusion from it.

Answers track the image whether or not the truth changes. On T-5 the models change their answer when the truth flips (mean CCR 0.241) almost as often as when it is held (mean FSR 0.216, 90% of CCR), and on T-2 the figures are 0.435 and 0.335. Little of either change is decoding noise. Shown the identical image twice, they disagree only 5.0% of the time on T-5 (28/559) and 3.9% on T-2. Part of it is render noise, though. The same geometry re-rendered with new noise changes 5.9% of Pixtral-12B's greedy answers and 4.0%–17.8% of the API models' (Table 25), so FSR mixes a response to the intervention's image change with a response to changes no person would notice.

Table 19: **Every cell on `merged_g9` and on `merged_r3`** (the plan's descriptive comparison). The two builds share questions, not pairs: `merged_r3` keeps 55 of `merged_g9`'s treatments with new controls, re-renders and re-answers every record and adds fresh items, so a row is not a paired comparison and neither column replicates the other. Bold: $p_{\text{room}} < 0.05$. Four cells are nominal on `merged_g9`, two on `merged_r3`, one on both. The sign of Δ (zero counted as a sign) differs in 11 of 22 cells and reverses in 6; Δ is lower on `merged_r3` in 11. Mean Δ over the cells of a template: T-2 $+0.061$ and $+0.099$; T-5 $+0.028$ and $+0.025$. †T-2 fails G1 on `merged_r3`: descriptive only, no conclusion under the pre-registered rule.

		`merged_g9`			`merged_r3`			
Model	**T**	n	Δ	p_{room}	n	Δ	p_{room}	change
Qwen2.5-VL-7B	T-2†	35	$+0.029$	1.000	26	$+0.000$	1.000	-0.029
	T-5	66	$+0.030$	0.781	52	$+0.058$	0.508	$+0.027$
Pixtral-12B	T-2†	35	-0.029	1.000	26	-0.038	1.000	-0.010
	T-5	66	-0.030	0.774	52	$+0.000$	1.000	$+0.030$
Gemini-3.1-Flash-Lite	T-2†	35	$+0.057$	0.809	26	$+0.192$	0.180	$+0.135$
	T-5	66	$+0.061$	0.529	52	$+0.000$	1.000	-0.061
Gemini-3-Flash	T-2†	35	$+0.057$	0.781	26	$+0.308$	0.063	$+0.251$
	T-5	66	$+0.182$	**0.031**	52	$+0.000$	1.000	-0.182
Gemini-3.5-Flash	T-2†	35	$+0.229$	**0.035**	26	$+0.115$	0.563	-0.113
	T-5	66	$+0.318$	**0.0026**	52	$+0.173$	**0.048**	-0.145
GPT-5.4-nano	T-2†	35	$+0.057$	0.766	26	-0.077	0.688	-0.134
	T-5	66	$+0.106$	0.092	52	$+0.058$	0.656	-0.048
GPT-5.4-mini	T-2†	35	$+0.029$	1.000	26	-0.038	1.000	-0.067
	T-5	66	-0.182	**0.034**	52	$+0.038$	0.625	$+0.220$
GPT-5.4	T-2†	35	$+0.086$	0.590	26	$+0.077$	0.781	-0.009
	T-5	66	-0.045	0.613	52	$+0.038$	0.781	$+0.084$
Claude-Haiku-4.5	T-2†	32	-0.156	0.125	23	$+0.130$	0.508	$+0.287$
	T-5	54	-0.074	0.388	40	$+0.025$	1.000	$+0.099$
Claude-Sonnet-5	T-2†	35	$+0.086$	0.637	26	$+0.154$	0.289	$+0.068$
	T-5	66	-0.061	0.596	51	-0.020	1.000	$+0.041$
Claude-Opus-5	T-2†	35	$+0.229$	0.056	26	$+0.269$	**0.039**	$+0.041$
	T-5	66	$+0.000$	1.000	52	-0.096	0.227	-0.096

H.3 Physical plausibility of the moved world-lines

Two defects of the generator's placement, found after the release was answered, are disclosed here with their sensitivity analyses (E24, E25). Neither is a gate (there are still nine), and no registered result is replaced. Every restriction below is descriptive.

The first (E24) leaves objects floating. The intervention moves only the target's own object, so whatever stands on or in it (a lamp, a book stack, a cup) stays where it was and hangs in mid-air in both moved images, S^{+} and S^{-}. The base image S is unaffected. The scene graph's recorded support relation marks only 24 of the affected records, so we detect them geometrically, flagging any object whose centre lies inside the target's horizontal box and whose base lies between 0.01 m above the target's base and 0.05 m above its top. Twenty-eight of the 78 released pairs are affected (56 of 156 records, 20 T-5 and 8 T-2 pairs), as is one of the 6 top-up items (bedroom_343) and 33 of the 101 `merged_g9` pairs D3 used. The same objects sit at the same place in both twins, so the defect cannot separate the arms.

Placements were also not checked against fixed furniture (E25). The collision test covered movable objects only, never tables, beds, sofas, shelves or walls. When no supporter was recorded, the support check accepted anything underneath, so §3's earlier "collide with nothing and rest on a surface" was never enforced. A new check flags a move that sinks the object at least 10 cm into another solid, pushes it at least 10 cm further out of its room's box, or stands it on too little support (Table 21). It flags 42 records in 38 pairs (20 by interpenetration, 19 by leaving the room, 1 by support, and a pair can meet more than one). One rater, an assistant model and not a person, classified every flagged record's renders: 19 visible, 11 subtle, 12 not visible, 0 false positives. 27 pairs are visible or subtle, on one record only in 24 (14 treatment, 10 control), and among the top-up items 2 are (diningroom_105, diningroom_117). The classification is descriptive and not a human rating. Only 31 of the 78 released pairs are clean of both defects.

Table 20: **Power of the primary test, and a bound on the pooled effect** (the release, `merged_r3`; not registered). Upper block: power of the room-level sign flip of Table 15 by simulation on the release's own rooms, 10,000 simulated data sets per entry at seed 20260920 (95% Monte Carlo half-width 0.010). A pair's d is $+1$ or -1 with probability $(p_{\text{disc}} \pm \Delta)/2$, p_{disc} the discordance $(b+c)/n$ pooled over the eleven models; *where $\Delta > p_{\text{disc}}$ the discordance is raised to Δ. Entries: $\alpha = 0.05$ / Bonferroni over the 22 cells ($\alpha = 0.0023$), two-sided. Pairs are independent within a room and every cell is given all of the template's pairs, so these powers are upper bounds; the ρ rows let the pairs of a room correlate at ρ. min. Δ: the smallest Δ (grid 0.01) with power $\geq 80\%$; rooms: the rooms needed for 80% after Bonferroni, pairs per room resampled from the release's. Middle block: Δ averaged over the models, each on its usable pairs, with a 95% interval and one-sided 95% bound from a joint room bootstrap (one resample of rooms for every model, 10,000 resamples at seed 20260920); room p: the sign flip on the per-room sums of d over the models. The single-model bound is the 95th percentile of each cell's own interval bootstrap (Table 15). Bottom block: CCR and FSR averaged over the models, and the disagreement between the two administrations of one byte-identical base render (Table 24), pooled over them. The drift anchor (Table 25), the same geometry re-rendered with new noise and answered again, changes 4.0–17.8% of the API models' answers and 5.9% of Pixtral-12B's (decoded greedily). †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

	T-2†	T-5
rooms / pairs	17 / 26	27 / 52
p_{disc}	86/283 = .304	134/559 = .240
power, $\Delta = 0.05$	.03 / .00	.06 / .00
power, $\Delta = 0.10$	.06 / .00	.19 / .02
power, $\Delta = 0.15$	.13 / .00	.46 / .08
power, $\Delta = 0.20$	.25 / .01	.81 / .28
power, $\Delta = 0.30$	.71 / .07	1.00 / .88*
$\Delta = 0.20$, $\rho = 0.2$	.23 / .01	.74 / .21
$\Delta = 0.20$, $\rho = 0.5$	.19 / .01	.62 / .13
min. Δ for 80%	.33* / .56*	.20 / .28*
rooms for 80%, Bonf., $\Delta = 0.10$	311 (≈476 pairs)	196 (≈377 pairs)
rooms for 80%, Bonf., $\Delta = 0.20$	77 (≈118 pairs)	49 (≈94 pairs)
pooled Δ, eleven models	$+0.099$ [$+0.025$, $+0.203$]	$+0.025$ [-0.023, $+0.078$]
one-sided 95% bound; room p	$+0.183$; .028	$+0.069$; .392
pooled Δ, ten, without Qwen	$+0.109$ [$+0.027$, $+0.223$]	$+0.022$ [-0.029, $+0.079$]
one-sided 95% bound; room p	$+0.201$; .028	$+0.070$; .506
largest single-model bound	$+0.474$ (Gemini-3-Flash)	$+0.295$ (Gemini-3.5-Flash)
mean CCR / mean FSR	.435 / .335	.241 / .216
same-image retest	11/283 = .039	28/559 = .050

A defect that fell more often on one arm could mimic or mask sensitivity. Read as a classifier of treatment against control, every definition of either defect gives an AUC between 0.423 and 0.548 on both templates, and exact sign tests on the pairs flagged on one record only give $p \geq 0.289$ (Table 21). Failing to reject does not establish balance, but nothing here suggests the defects favour either arm.

Restricting the registered analyses to the pairs clean of a defect reverses no conclusion, though without E25's defects D3 is no longer significant after Holm (below). No cell of the 22-cell family is nominal on the pairs without floating objects (0), on those without a visible or subtle placement defect (0) or on those clean of both (0). The pooled Δ over the eleven models is $+0.025$ on T-5 with all pairs, $+0.026$ without floating objects and $+0.049$ without visible placement defects (on the affected pairs -0.009). The top-up test without its floating-object item has $n = 5$, 3/2/0 and $p = 1/27$, again its design's floor. The registered test stays $n = 6$, $p = 1/81$. D3's primary contrast holds on the pairs without floating objects ($D = -60$ over 662, $p = 0.0176$, rejected after Holm, 10 of 10 models lower) and is not detected on the affected ones ($D = 2$, $p = 0.911$). Without E25's defects its direction holds and its size grows ($D/n = -0.080$, and -0.101 clean of both, against -0.059 on all pairs), but p is 0.036–0.045 over the restrictions we ran, and none survives Holm. D3's pairs have no visual classification, so there a pair counts as affected by E25 when the machine flags its treatment or any of its controls. The image ablation is not recomputed on any subset. Its three arms show only the base image, which neither defect touches, so no restriction of it would correct anything.

The generator now has opt-in options that carry supported objects with the target and reject placements that interpenetrate fixed furniture or walls or lack support. They are off by default, so the release rebuilds byte for byte. A rebuilt release with them on, checked by the same two scripts, will be released after review.

Table 21: **Physical plausibility of the moved world lines** (the release, `merged_r3`; audit entries E24 and E25). *E24*: the generator moves only the target's own object, so what stands on or in it stays where it was and floats in both moved images (S^+ and S^-), on both records of a pair alike. *E25*: placements were tested against movable objects only; the machine flag marks a move that sinks the object at least 10 cm into another solid, pushes it at least 10 cm further out of its room's box (walls are not in the scene graph) or stands it off the floor on less than 50% of its base, over no supporter, or, tall and narrow, on too small a top (42 of 156 records, 38 pairs). Every flagged record's renders were classified by a single rater: 19 visible, 11 subtle, 12 not visible, no false positives (27 pairs visible or subtle, 24 of them on one record only); the classification is descriptive. Pairs: pairs with a flagged record; T/C: pairs flagged on the treatment record only / the control record only; AUC: how well the flag tells a pair's treatment record from its control, $0.5 + (p_T - p_C)/2$; p: exact two-sided sign test on the one-sided pairs. Lower block, descriptive: the registered analyses restricted to the pairs clean of, or affected by, a defect (restrictions, not corrections): pairs T-2 / T-5; nominal: cells of the 22-cell family with room-level $p < 0.05$; Δ pooled over the eleven models; the fresh T-2 test, items left: p (each at its design's floor); D3's sample-and-filter contrast (Table 7), D/n and room p, ^{h}Holm rejects. D3's base (`merged_g9`) and arm controls have no visual classification, so there a pair counts as affected by E25 when the machine flags its released treatment, its control or any arm's control. No gate was added and no registered result is replaced. The image ablation is not recomputed: its three arms show only the base image S, which neither defect touches. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

	T-2†				T-5			
Defect	pairs	T/C	AUC	p	pairs	T/C	AUC	p
E24: object left floating	8/26	0/0	.500	1.00	20/52	0/0	.500	1.00
E25: machine flag	9/26	2/6	.423	.289	29/52	14/12	.519	.845
E25: visible or subtle	5/26	2/3	.481	1.00	22/52	12/7	.548	.359
E25: visible	2/26	1/1	.500	1.00	16/52	9/6	.529	.607
E24 or E25 machine flag	14/26	1/4	.442	.375	37/52	7/8	.490	1.00
E24 or E25 visible/subtle	13/26	2/3	.481	1.00	34/52	7/5	.519	.774

			Δ, 11 models		fresh T-2	D3: SF vs. EX		
Pairs kept	pairs	nominal	T-2†	T-5	n: p	pairs	D/n	p
all pairs	26 / 52	2	+.099	+.025	6: 1/81	101	−.059	.019^h
no E24	18 / 32	0	+.082	+.026	5: 1/27	68	−.091	.018^h
E24 affected	8 / 20	0	+.136	+.023	1: 1/3	33	+.006	.911
no E25 visible/subtle	21 / 30	0	+.100	+.049	4: 1/9	49	−.080	.036
E25 visible/subtle	5 / 22	0	+.093	−.009	2: 1/9	52	−.039	.107
neither defect	13 / 18	0	+.078	+.036	3: 1/3	36	−.101	.045
either defect	13 / 34	0	+.120	+.019	3: 1/27	65	−.036	.060

I AUDIT TRAIL

Table 22 lists the twenty-four registered defects in the order we found them, E1–E13 before the rebuild and E14–E25 during it, with E23 recorded in the repository only and not counted. Table 23 gives the headline numbers on each build.

E1. G5 caught a T-5 text leak. An early T-5 build reached a blind baseline of 0.656 through the category-pair arm alone, meaning the two object categories a question names predicted its answer 65.6% of the time. Model behaviour had made us suspicious, but it cannot separate a weak model from a text-solvable item set. A baseline running no model can. The build was rejected and the question generator changed.

E2. Referential ambiguity, which a blind probe caught and no gate did. On the build before the repair, the named objects in 3 of the 11 T-5 items given to the probe could not be uniquely identified

Table 22: **Audit trail in full.** Twenty-four registered defects: two caught by a gate, one by another executable check (the pre-registered render calibration), fourteen by our own re-measurement or review, two by something outside our process (a blind answerability probe and a model's own refusal to answer), and five by external review, each raised by a human reader and, independently, in an LLM-based review of a draft (E11, E13–E16). ^{u}A claim we had written that re-measurement did not support (eight: E3, E7, E10, E14, E18, E19, E20, E25). E23, in italics, is recorded in the repository and not counted. Each entry is described below.

	Defect	Caught by	Consequence		Defect	Caught by	Consequence
E1	T-5 text leak	G5 blind baseline	build rejected	E14^u	G6 blind to on-screen motion	external review	G6 extended; release rebuilt
E2	Referential ambiguity	blind probe, *not* a gate	3/11 items unidentifiable	E15	BH misstated in a caption	external review	caption corrected
E3^u	"no image signal"	re-measurement	withdrawn	E16	Test's execution history undisclosed	external review	history released
E4	CCR_{dir} vs. FSR	metric review	retracted	E17	T-5 enlarged after a null	our review	disclosed; T-5 rests on the release
E5	Pseudo-replication	de-duplication	n inflated about $2\times$	E18^u	Renders reproduce on one CPU type only	pre-registered calibration	re-rendered, re-answered
E6	False replication	overlap check	22/47 questions shared	E19^u	Retest understated volatility	pre-registered drift anchor	claim qualified
E7^u	Same, again	independent re-audit	withdrawn	E20^u	Two headline results fail on the release	re-analysis	removed from headline
E8	Near-black renders	a model's refusal, *not* a gate	G8 added; 8 pairs dropped	E21	Plan's rules conflict on T-2	at analysis	conflict stated
E9	Refusals parsed as answers	independent re-derivation	parser fixed	E22	Rotation path skipped a check	generator review	check added; pair re-solved
E10^u	De-dup rule implemented twice	independent re-derivation	accuracy corrected	*E23*	*Tooling defaults failed silently*	*planning audit*	*fixed; repository only*
E11	Same-room pairs treated as independent	external review	BH survivors $2\to0$	E24	Objects left floating	our review	disclosed; sensitivity
E12	Phrases rebind after the move	G9, re-implemented	40 of 141 pairs dropped	E25^u	Placement unchecked against fixed furniture	our review	disclosed; sensitivity
E13	Gate chronology overstated	external review	claims rewritten				

from the render (Appendix C). No gate looked at referring expressions then. The repair became a construction rule, and G9 later a gate.

E3. "T-5 carries no image signal", withdrawn. On run7, a build carrying the E2 defect, the real and mismatched arms tied, and we recorded that T-5 carries no image signal. The log then read real = mismatched = 20/37 (0.541), p = 0.2054, each arm over its own parsed items, and the current parser gives 19, 20 and 15 of 36 (real, mismatched, grey), $p = 0.205$. The design was underpowered, and the data also carried a known defect. A null on defective data is not evidence of absence, and we withdrew the claim without putting its opposite in its place. The registered test gave $p = 0.023$ on `merged_g9`, a build enlarged after a null (E17), and on the release T-5 fails to show answerability ($p = 0.121$, $n = 51$, E20). A null on defective data is not evidence of absence, and a positive on a build grown until significant is not evidence of presence.

E4. CCR_{dir} compared against FSR, retracted. CCR_{dir} requires both sides exactly correct, while FSR asks only whether the answer changed, and the difference of two quantities with different criteria is not a sensitivity measure. Metric review caught it, not a gate.

E5. Pseudo-replication, caught by de-duplication. Records sharing a (scene, question) were being counted as independent items. On the release that would inflate n by a factor of 2.05. De-duplication by (scene, question) is now part of the registered analysis.

E6 and E7. False replication, withdrawn twice. We twice described a rebuilt dataset as an independent replication when it overlapped the earlier one (E6) or contained it (E7). Construction and rendering are deterministic, so rebuilding from the same code on overlapping scenes produces identical records. In E6, run5 and run6 share 22 of 47 (scene, question) stems, 21 of them on the same base camera. In

E7, all 120 of run11's records recur unchanged in `merged`, so the "disjoint replication at $p = 0.039$, $n = 35$" we had recorded was the same data. An overlap check on keys, truths and image hashes caught both. We now run that check before any replication claim, and it ships with the earlier builds' data. The same rule applies to the release. Of its 156 records, 56 match a record of `merged_g9` in question, truth and geometry, and 70 of its 76 questions were asked there, so it is not an independent replication of `merged_g9`. The only form of replication this project can honestly offer is cross-model replication on one dataset.

E8. Near-black renders, caught by a model's refusal and by no gate. Claude-Haiku-4.5 said 61 times that it could not see the objects named, and Claude-Sonnet-5 described an image as "almost entirely black". It was, and so were 32 of 596 renders, one scene at 0.11% non-black pixels. Seven gates checked counts, balance, displacement, blind baselines, arm separability and margin, but none checked that the picture had anything in it. We added G8, dropped 8 pairs and recomputed every number. At that step the ablation strengthened on both templates, T-2 from 3.5×10^{-4} to 2.6×10^{-4} and T-5 from 0.0091 to 0.0078 (Table 23), because near-black items had been diluting the real arm toward the grey one.

E9. Refusals parsed as answers, and a correct answer discarded. An independent re-derivation of every published number found both. Appendix E gives the two defects, the rules that now correct them and their effect. An earlier draft disclosed them without correcting them, reasoning that the results had been seen. A pre-registration does not oblige anyone to keep a confirmed implementation error, and no conclusion moves.

E10. One de-duplication rule, two implementations. The registered rule claims the (scene, question) key on the first record and drops the question entirely if that record's answer is unusable. The script that reproduces Table 3 claimed the key only after the usability check, so a pair partner silently substituted. The partner's base render is the same image, so the substitute was that image answered a second time, and one accuracy figure mixed two administrations. The figure-drawing script implemented the registered rule correctly, so the two disagreed in print. Both now call one function. This is the failure the paper argues against, a criterion implemented twice and inconsistently, occurring inside the paper itself.

E11. Pairs within a room treated as independent, found by external review. We tested each cell with the record-level McNemar test and then required a room-bootstrap interval to exclude zero. The second step does not repair the first, and in practice it removed nothing. With rooms as the unit of inference, the 2 cells that had survived Benjamini–Hochberg on `merged_g8` no longer do (Appendix H).

E12. Phrases rebinding after the intervention, caught by G9 once it applied the generator's rule. G9 was written because an external reader pointed out that moving one object can change what "the leftmost chair" denotes. Its first implementation ranked projected object centres, because the segmentations the generator used had not been kept. It judged only the treatment world-line and counted a pair as exposed only when both members of an ordinal group were named, and it passed the nine groups it could judge. Re-running the builder's own ray-cast from the original scenes and judging every phrase on both world-lines failed 40 of 141 pairs. They were dropped, and everything was recomputed on the 101 that remained (`merged_g9`). On the rebuild's set G9 caught 3 more rebinds, all on the control. A gate that checks a proxy for the rule can pass while the rule fails.

E13. Gate chronology overstated, found by external review. An earlier abstract said nine gates "inspect the data before any model is run". Five did. The other four were written after models had answered the build, one of them prompted by a model's refusal. All nine read only data, and the earlier sentence conflated that with being fixed in advance. G6 was amended again on 24 September, after models had answered `merged_g9` (E14). With that amendment, all nine were fixed before any model answered the release, and ran on it after one model had answered the assembled records (Appendix A). §4 and Appendix A date each gate.

E14. G6 blind to on-screen motion, found by external review. G6 read pixel RMSE and changed-pixel fraction, and `merged_g9` passed it. An external reader pointed out that G6 did not measure the moved object's travel across the image. On `merged_g9` the exact controls travelled further on screen than their treatments in 73 of 101 pairs (room-level sign flip $p = 4.1 \times 10^{-5}$), and a responder reading screen travel separated the arms (Table 6). Our earlier text had called the arms indistinguishable in image space. G6 gained the screen statistic on 24 September, and `merged_g9` fails it (AUC 0.727

on T-2, 0.632 on T-5). We re-solved the controls to match the treatment's on-screen travel within 10% (Appendix B.3) and rebuilt the release, where the AUCs are 0.501 and 0.500. The rebuild cost T-2 its G1 pass.

E15. Benjamini–Hochberg misstated in a caption, found by external review. A caption said that Benjamini–Hochberg could reject only if the smallest p were at most $0.05/22$. Benjamini–Hochberg is a step-up procedure that rejects the k smallest p-values for the largest k with $p_{(k)} \leq k \cdot 0.05/22$. The error was in the caption's explanation and not in the computation, so no result changes. The caption of Table 15 now states the rule and the closest miss.

E16. The registered test's undisclosed execution history, found by external review. The registered T-2 ablation had been applied to every T-2 build from run5 on, and the paper reported only the latest, on a build revised after models had answered earlier ones. That is the registered test applied post hoc and presented as if it confirmed. The earlier builds' ablation data are now released, Tables 2 and 10 list every application, and each result is worded for the build it comes from.

E17. T-5 enlarged after a null, found in our review of an external point. After run11's T-5 ablation came back at $p = 0.0801$ ($n = 33$), T-5 was enlarged by merging builds until it reached $p = 0.0091$ ($n = 95$). T-5 was never registered, and collecting further after a null inflates the error rate. We disclose it as optional continuation (Appendix D), and the T-5 claim now rests on the release alone.

E18. Renders reproduce only on the same CPU architecture, caught by the pre-registered calibration. The release plan required released controls re-rendered on the new host to match their released renders pixel for pixel. The first one kept its geometry and mean brightness (91.36 against 91.35) but not its noise, with 66% of pixels differing by a median of 1 and at most 198 levels. On the same host, 4 and 8 threads rendered it identically. Only 16 of the rebuild's 516 files match an earlier render once metadata are stripped. Our earlier text said the renders reproduce pixel for pixel, which holds only on the same CPU architecture. As the plan's contingency required, the whole release was rendered on one host and answered again by every model (Appendix D).

E19. The retest bound understated answer volatility, caught by the pre-registered drift anchor. We had read the 4.6% disagreement between two administrations of one byte-identical image as the models' volatility. The drift anchor re-rendered `merged_g9`'s exact controls with identical geometry and new noise, which changed 6 of 101 of Pixtral-12B's greedy answers. That model never changes its answer on an identical image (its retest rate is zero), so the change comes from the render. The API models changed 4.0%–17.8% of their answers, which may also include sampling and version drift, and every model but Gemini-3.1-Flash-Lite exceeds the plan's 5.9% retest bound (Table 25). Part of what models change is render noise, and the volatility claim is qualified.

E20. Two headline results do not hold on the release. The T-5 ablation, $p = 0.023$ on `merged_g9`, is 0.121 on the release. On the release's 51 T-5 questions, `merged_g9`'s own answers already gave 0.086, so the change comes largely from which questions the rebuild kept. The potted plant's across-model excess, selection-adjusted $p = 0.0015$ on `merged_g9`, has $p = 0.068$ on the release. Both are removed from the headline. The first is reported as a failure to show answerability at $n = 51$ and the second as exploratory.

E21. Two rules of the plan conflict on T-2, found at analysis. The release plan names the top-up T-2 ablation its confirmatory test, and the same plan gives no conclusion for a template that fails G1, which T-2 does on the release. We report the test as the plan's confirmatory test and state the conflict. The test has $n = 6$, far below G1's 30, and sits at its design floor, $p = 1/81$ (Appendix D).

E22. A review of the generator found that the rotation path skipped the distance-convention check. The treatment path and the floor path of the control construction checked that a second distance reading gives the same label, but the rotation path did not. On `merged_g9`, 1 released control fails the check. The rotation path now applies it, the pair was re-solved for the release, and no released record fails (Appendix E.1).

E23. Tooling defaults, recorded in the repository only. While planning the rebuild we found analysis tools whose defaults pointed at earlier builds or failed without an error. G9 defaulted to the old segmentations, G7 passed at zero coverage when run from another directory, a merge warning never fired, and one analysis crashed on single-sided arms. All were fixed before any released number was computed. No published number was affected, so the entry is not counted.

E24. Objects left floating, found in our own review. While choosing the example pair of Figure 1 we saw a lamp hanging in mid-air. The generator moves only the target, so whatever stands on it stays behind, in 28 of the 78 released pairs and in earlier builds (33 of the 101 pairs of `merged_g9`). The defect is identical in the two twins. We disclose it with a sensitivity analysis on the unaffected pairs, which changes no conclusion (Appendix H.3).

E25. Placements never checked against fixed furniture, found while auditing E24. Our text said a placed object "collides with nothing and rests on a surface", but the generator tested collisions with movable objects only and accepted any object underneath as support. On inspection, 27 of the 78 released pairs show interpenetration or implausible support, 24 of them on one record only. The defects do not detectably separate the arms, and restricting to the clean pairs reverses no conclusion, though D3 is then not significant after Holm (Appendix H.3). The generator now has the checks, off by default, for a rebuild after review.

Table 23: **The four builds side by side.** G8 (render content) drops the pairs with a render that carries no visible content, G9 (stable referential binding) those whose ordinals do not denote the same objects on both world-lines; `merged_g8` and `merged_g9` are each a subset of the build to their left; `merged_r3` is not a subset of `merged_g9` but a recomposition (re-solve, re-render, re-answer, top-up, G9): 55 of its 78 pair ids recur, re-rendered and answered again. All four are scored with the current answer parser. G1 needs at least 30 distinct questions per template; ✗ marks a failing value. G6's screen statistic (the moved object's on-screen displacement) was added for the rebuild; on `merged`, `merged_g8` and `merged_g9` it was not in force and is shown for reference (^{s}), and on `merged` it cannot be computed (no scene graph for the moved object of 16 records). *Gates failed* lists the gates in force a template fails on that build; † marks the results of such a template, descriptive only under the pre-registered rule. The ablation p is the one-sided exact trend test (real $>$ mismatched $>$ grey, Gemini-3-Flash; registered for T-2), questions de-duplicated by (scene, question), n in parentheses. Cell counts are over the 22-cell family, as record-level McNemar / room-level sign-flip; *nominal* is $p < 0.05$ uncorrected, BH at FDR 0.05, Bonferroni at 0.05. Degeneracy flags use the room-level rule of Appendix G.

	`merged` (before G8)	`merged_g8` (after G8)	`merged_g9` (after G9; previous release)	`merged_r3` (rebuild; released)
records	298	282	202	156
matched pairs, T-2 / T-5	52 / 97	48 / 93	35 / 66	26 / 52
scenes	38	35	34	32
G1: distinct questions, T-2 / T-5	45 / 95	42 / 91	31 / 65	25✗ / 51
G6: screen-motion AUC, T-2	–	0.746✗^{s}	0.727✗^{s}	0.501
G6: screen-motion AUC, T-5	–	0.626✗^{s}	0.632✗^{s}	0.500
gates failed, T-2 / T-5	G8, G9 / G8, G9	G9 / G9	– / –	G1 / –
ablation trend p, T-2	3.5×10^{-4} (44)†	2.6×10^{-4} (41)†	4.4×10^{-5} (30)	1.8×10^{-4} (25)†
ablation trend p, T-5	0.0091 (95)†	0.0078 (91)†	0.023 (65)	0.121 (51)
nominal cells, record / room	4 / 3	4 / 2	3 / 4	1 / 2
BH survivors, record / room	0 / 0	2 / 0	1 / 0	0 / 0
Bonferroni survivors, record / room	0 / 0	0 / 0	1 / 0	0 / 0
degeneracy-flagged cells	4/22	4/22	4/22	3/22

J COVERAGE AND MODEL ACCESS

Table 24 gives every denominator. Claude-Haiku-4.5 has 6 unusable answers on T-2 and 18 on T-5, and 23 of those 24 are statements that it cannot see the objects. That leaves 23 and 40 usable pairs. Claude-Sonnet-5 has 1 unusable T-5 answer, a reply the parser cannot assign to either option. Every other model's answers are all usable. $\mathrm{CCR}_{\mathrm{dir}}$, the rate at which both sides of a treatment pair are exactly right, peaks at 0.423 on T-2 (Gemini-3-Flash, 11/26) and 0.327 on T-5 (Gemini-3.5-Flash, 17/52), against rates of $1/16$ and $1/4$ under independent uniform guessing.

Refusals. On the unfiltered `merged` build, Claude-Haiku-4.5's 61 cannot-see replies led us to its near-black renders (E8). On the release every render passes G8, yet the model still gives 23 such replies, so its abstention does not track a defect any gate detects. Rule R3 scores them unusable.

Test–retest. The two records of a pair share one byte-identical base render, so each base image is answered twice with the same question. Disagreement is at most 9/52 (Claude-Opus-5 T-5, one-sided

95% Clopper–Pearson upper bound 0.283), 4.6% pooled over models (39/842, bound 6.0%), and zero in 7 of the 22 cells. The bound is the exact binomial one for the count, as if every administration were independent. Administrations cluster by room and by model, so the bound is too narrow if disagreements cluster. It covers base images only.

Render noise (E19). Table 25 anchors that rate against a different source of change: the same geometry, re-rendered with new noise and answered again. Pixtral-12B, decoded greedily on a local machine, changed 6 of 101 answers, so a new noise realisation alone moves a deterministic model. The API models changed 4.0%–17.8%, which also includes their own sampling and any change of the model behind its name between the two sessions.

Access. Every answer on the release, and every answer on a D3 arm, was collected on 24–25 September 2026. Pixtral-12B ran locally in bfloat16 on Apple MPS and Qwen2.5-VL-7B on a rented remote GPU, both with greedy decoding and short output budgets, ample for the one-phrase answers the prompt asks for. Gemini-3.1-Flash-Lite, Gemini-3-Flash, Gemini-3.5-Flash, GPT-5.4-nano, GPT-5.4-mini and GPT-5.4 were queried at temperature 0, with Gemini's thinking left at its default. Of Claude Haiku 4.5, Sonnet 5 and Opus 5, the latter two ran at low reasoning effort (Haiku 4.5 exposes no effort setting). API output budgets were 4096 tokens. These are the settings of the previous release, when the API models were queried on 2–6 September 2026. The Claude API rejects a temperature parameter, so that family's decoding is not pinned as the others' is, and we make no claim that its answers are deterministic. D3's arms were answered on the intervened render only, and the ablation's mismatched and grey arms on the base render only. Responses that failed with a server-overload error were re-queried, as were twelve Gemini-3.5-Flash responses truncated on the previous release under an earlier 1024-token budget. The criterion was the failure, never the answer. No server-side model fallback was enabled on any vendor, because a silent substitution would replace the model under test without changing the name we report.

Table 24: **Denominators, unusable answers, the test–retest control and strict accuracy.** *treat*/*ctrl* count *records* of each arm whose two answers are both usable; *unusable* counts individual answers. CCR_{any}, FSR and Δ (Table 15) use the intersection, the n matched *pairs* whose two records are both usable, and so do the four columns on the right. *acc* is de-duplicated by (scene, question) over the first record, which is dropped entirely when its answer is unusable rather than letting the pair partner substitute (E10). *retest* is disagreement between the two administrations of the same byte-identical base render. CCR_{dir}: the treatment record is right on both the base and the intervened render; $ctrl_{both}$: the control record is right on both; *all3*: the treatment base, the treatment intervened and the control intervened answers are all right (S, S^+, S^-). Under independent uniform guessing the two-answer columns have chance $1/16$ on T-2 and $1/4$ on T-5, and *all3* $1/64$ and $1/8$. Unusable answers occur only for Claude-Haiku-4.5 and Claude-Sonnet-5; for every other model *treat* = *ctrl* = n. †T-2 fails G1 on this build: descriptive only, no conclusion under the pre-registered rule.

Model	**T**	treat	ctrl	unusable	n	acc	retest	CCR_{dir}	$ctrl_{both}$	all3
Qwen2.5-VL-7B	T-2†	26	26	0	26	6/25	0/26	0/26	6/26	0/26
	T-5	52	52	0	52	29/51	0/52	9/52	26/52	6/52
Pixtral-12B	T-2†	26	26	0	26	8/25	0/26	1/26	7/26	1/26
	T-5	52	52	0	52	25/51	0/52	5/52	23/52	3/52
Gemini-3.1-Flash-Lite	T-2†	26	26	0	26	15/25	0/26	8/26	6/26	6/26
	T-5	52	52	0	52	28/51	0/52	10/52	26/52	7/52
Gemini-3-Flash	T-2†	26	26	0	26	16/25	1/26	11/26	10/26	8/26
	T-5	52	52	0	52	32/51	2/52	13/52	23/52	8/52
Gemini-3.5-Flash	T-2†	26	26	0	26	16/25	1/26	10/26	9/26	6/26
	T-5	52	52	0	52	31/51	1/52	17/52	26/52	12/52
GPT-5.4-nano	T-2†	26	26	0	26	6/25	1/26	1/26	3/26	0/26
	T-5	52	52	0	52	31/51	1/52	10/52	27/52	6/52
GPT-5.4-mini	T-2†	26	26	0	26	10/25	1/26	3/26	7/26	0/26
	T-5	52	52	0	52	29/51	6/52	5/52	28/52	3/52
GPT-5.4	T-2†	26	26	0	26	14/25	0/26	10/26	5/26	5/26
	T-5	52	52	0	52	28/51	2/52	10/52	23/52	6/52
Claude-Haiku-4.5	T-2†	23	24	6	23	7/25	3/23	2/23	5/23	2/23
	T-5	42	46	18	40	24/47	4/40	5/40	20/40	2/40
Claude-Sonnet-5	T-2†	26	26	0	26	14/25	3/26	5/26	6/26	4/26
	T-5	51	52	1	51	29/50	3/51	9/51	24/51	4/51
Claude-Opus-5	T-2†	26	26	0	26	14/25	1/26	6/26	8/26	3/26
	T-5	52	52	0	52	25/51	9/52	6/52	19/52	2/52

Table 25: **The drift anchor.** Per model, how often its answer from the previous release on a released exact control of `merged_g9` differs from its new answer on the same control re-rendered in the D3 session (EX): the same geometry, answered again. Six of the 101 re-renders are byte-identical to the previous release's render (metadata stripped); the others carry a new noise realisation. The bound is the plan's same-image retest bound, 5.9% (typed into analyze_r3 from the plan; no project function computes it); the pooled one-sided 95% Clopper–Pearson bound of the test–retest rate is 5.7% on `merged_g9` (50/1096) and 6.0% on `merged_r3` (39/842). [l]Decoded greedily on a local machine, so on an identical image it must repeat itself; the API models add their own sampling and possible version drift (4.0–17.8%). Re-rendering one released control on the release's rendering host, a server with a different CPU model, changed 66% of its pixels (median absolute difference 1, maximum 198 of 255); the same host at 4 and 8 threads rendered it identically.

Model	n	changed	%	identical re-render	new noise	$>5.9\%$
Pixtral-12B[l]	101	6	5.9	0/6	6/95	✓
Gemini-3.1-Flash-Lite	101	4	4.0	0/6	4/95	
Gemini-3-Flash	101	12	11.9	0/6	12/95	✓
Gemini-3.5-Flash	101	11	10.9	0/6	11/95	✓
GPT-5.4-nano	101	17	16.8	1/6	16/95	✓
GPT-5.4-mini	101	14	13.9	1/6	13/95	✓
GPT-5.4	101	17	16.8	1/6	16/95	✓
Claude-Haiku-4.5	91	13	14.3	2/5	11/86	✓
Claude-Sonnet-5	101	14	13.9	0/6	14/95	✓
Claude-Opus-5	101	18	17.8	0/6	18/95	✓