ESCBENCH: BENCHMARKING CONTEXT-SENSITIVE SPATIAL DEIXIS IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating spatial demonstratives requires asking whether a response to context retains its direction and distance profile across observations of the same situation. Recognition accuracy, prompt stability, and aggregate human resemblance each leave this question unresolved. We introduce ESCBench, a diagnostic benchmark of English *this* and *that* that crosses changes in depicted context with changes in observation under a fixed speaker task. Its 576 tabletop scenes vary distance, tool configuration, interlocutor layout, shape, and color, yielding 2,304 images from four viewpoints. Linked comparison identities support tool and layout contrasts within each view; seven protocols connect these contrasts to attribute recognition and elicitation sensitivity. Across 22 vision–language model deployments, the same tool contrast can reverse across views despite high overall recognition and low overall task-form sensitivity. Similar external-human distribution scores can also accompany a large positive tool response or a near-zero one, while pooling views can yield a closer match than every individual view. These findings distinguish contextual responses that agree across observations from observation-dependent responses and aggregate resemblance produced by mixing them. ESCBench makes these distinctions directly testable through matched comparisons.

1 INTRODUCTION

Spatial reference relates objects to speakers and addressees, a requirement for agents that coordinate with human partners (Ichter et al., 2023; Chang et al., 2025). English spatial demonstratives, *this* and *that*, offer a focused test: human use varies with distance, action possibilities, and speaker–addressee configuration (Coventry et al., 2008; 2023). With one target and two candidate words, a vision–language model (VLM) must express a relation to the designated speaker from the available image.

The central question is whether a contextual response survives a change in observation. A model may identify an object correctly, respond consistently to prompt variants, and resemble a human distribution, yet reverse its response to a tool or interlocutor configuration when the camera changes. To determine whether the same contextual contrast retains its direction and distance profile, the situation change must be repeated within each observation view while the speaker task remains fixed.

Existing studies establish gaps between vocabulary and perspectival-word use, including human comparisons, cue ablations, and prompt variants (Dong et al., 2026), and evaluate multilingual demonstratives using distance-conditioned human distributions (Watanabe et al., 2026). SCOPE tests spatial consistency across perspectives in both single-view and multi-view settings (Kim et al., 2026). We extend this diagnosis to a fixed speaker's tool and layout responses, linking each contextual contrast across distance and observation.

We introduce ESCBench, a controlled tabletop benchmark whose construction, visual comparisons, and evaluation dimensions are summarized in Figure 1. Green-and-yellow MIMo (Mattern et al., 2023) is the speaker; black-and-white iCub is the addressee, beside MIMo or across the table. MIMo points with a bare hand or held rod at one object. Crossing twelve distances, two tools, two layouts, three shapes, and four colors yields 576 scenes. Speaker-eye, addressee-eye, external (Exo), and top-down (Top) cameras provide 2,304 separately queried images. Matched tool and layout pairs are evaluated within each view, then compared across views.

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

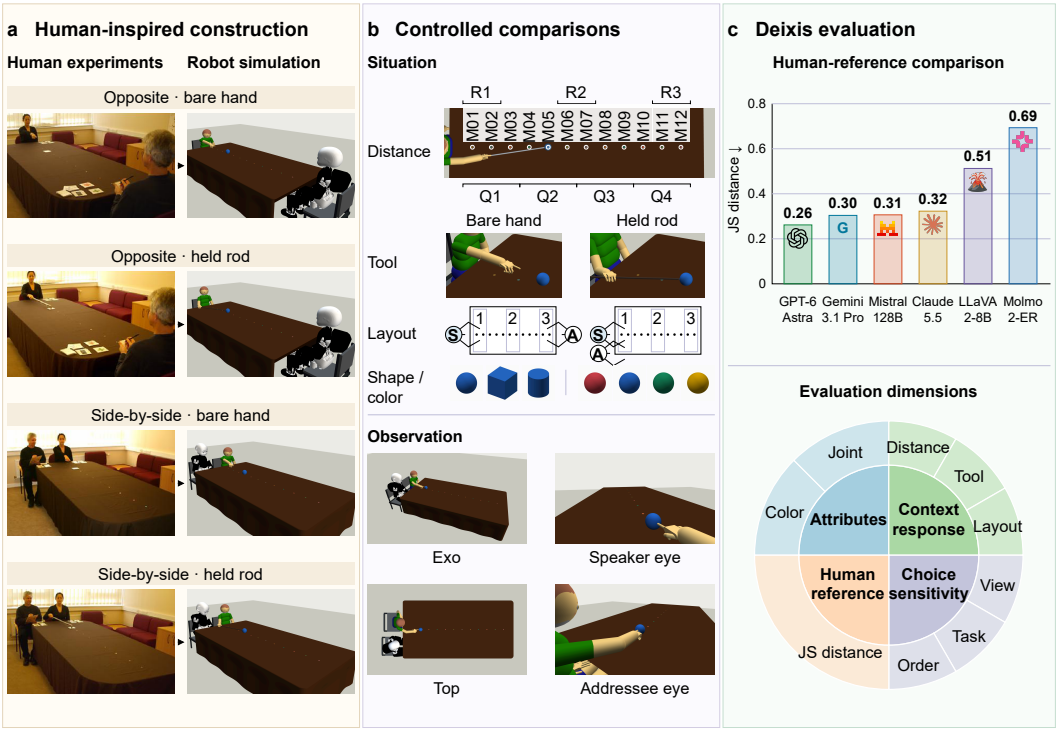


Figure 1: **From human studies to controlled spatial-deixis evaluation.** (a) Four human experimental conditions (Coventry et al., 2008), crossing tool use and interlocutor layout, paired with robot presentations at M04. (b) Situation factors and four observations of one M02, bare-hand, side-by-side scene, with the speaker fixed. The distance strip shows alternative target positions M01–M12 (0.15–1.80 m in 0.15-m steps), not simultaneous objects. Q1–Q4 and R1–R3 are separate human-reference mappings (Coventry et al., 2008; 2023), not reachability regions. Tool closeups compare bare-hand and held-rod postures at M03 using the same camera and crop. Presentations and tool closeups are illustrative, not model inputs. Layout schematics mark speaker (S) and addressee (A); shape/color examples use simulation renders. (c) Six of 22 deployments: mean condition-wise base-2 Jensen–Shannon distance to six English region–layout distributions (Coventry et al., 2023) (T2/bare/Exo; lower is closer). Each condition uses the square root of JS divergence. This measures external resemblance, not current-image accuracy. The nested diagram organizes Table 1's four groups and nine measures, including task/order sensitivity; sector areas are schematic, not proportions or weights.

Seven protocols connect this design to recognition and elicitation checks. T1 requests a demonstrative; T2 adds color and shape. Further protocols request attributes alone, vary contextual wording, or reverse the candidate order on the same image. Option-order sensitivity is established in multiple-choice evaluation (Pezeshkpour & Hruschka, 2024); here it supplies a complementary check on the elicitation used to measure demonstrative choices. The analyses cover recognition and stability (RQ1), contextual responses across views (RQ2), and external human resemblance (RQ3).

Results from 22 deployments show how these comparisons change the interpretation (Table 1). Gemini combines perfect overall attribute recognition with 5.64% overall T1–T2 changes, yet its M04 tool contrast switches from $+12.50$ pp in Exo to -33.33 pp in Top (Figure 3a). Thus good recognition and modest task-form sensitivity coexist with an observation-dependent tool response. Qwen3.8-27B's pooled-view human-reference JSD is 0.133, below every single-view value (0.231–0.400; Figure 4e): the improved aggregate match does not characterize any individual observation. Conversely, Molmo2-ER's constant outputs explain its apparent stability without contextual responsiveness.

Contributions. (1) **Matched context-by-observation comparisons:** linked scene and image identities make the direction and distance profile of tool and layout responses comparable across views under a fixed speaker task. (2) **An integrated diagnostic resource:** seven protocols connect these comparisons to scored attributes, elicitation changes, and external human distributions, with explicit coverage and paired transitions. (3) **Empirical distinctions among superficially similar behaviors:** the analyses separate constant-output stability, observation-dependent contextual responses, and aggregate human resemblance that conceals conditional differences.

We will publicly release the stimulus images, annotations, prompts, and evaluation code upon acceptance.

2 Related Work

Spatial grounding and evaluation. Winoground, ARO, and What'sUp distinguish entity recognition from compositional or relational understanding (Thrush et al., 2022; Yuksekgonul et al., 2023; Kamath et al., 2023). SpatialVLM, SpatialRGPT, and RoboSpatial strengthen spatial grounding through data and representations (Chen et al., 2024; Cheng et al., 2024; Song et al., 2025). COMFORT evaluates reference-frame ambiguity and consistency; SPHERE organizes spatial skills hierarchically (Zhang et al., 2025b;a). SCOPE evaluates single-view spatial consistency and updating alongside multi-view integration across egocentric and allocentric tasks (Kim et al., 2026); MMSI-Bench focuses on reasoning across images (Yang et al., 2026). ESCBench examines a different response: whether changing a tool or interlocutor configuration changes a designated speaker's demonstrative choice in the same direction and distance range across observations.

Situated reference and demonstratives. PUB evaluates pragmatic interpretation, and RefOI studies referring expressions and human pragmatic preferences (Sravanthi et al., 2024; Ma et al., 2025). Dong et al. (2026) compare vocabulary, possessives, and demonstratives in humans and models, with prompt variants, cue ablations, and object/speaker identification checks. Watanabe et al. (2026) evaluate four-language demonstratives with three-word outputs, distance conditions, human distributions, and recognition-conditioned analysis. Building on these findings, ESCBench repeats matched tool and layout contrasts across distance and camera view. This reveals whether an apparently stable or human-resembling response preserves the same conditional pattern across observations. Attribute checks and external distribution distances support that interpretation within a common expected-image accounting; they do not supply current-image demonstrative labels.

Behavioral measurement. CheckList motivates capability-specific tests (Ribeiro et al., 2020); prompt-format and multi-prompt studies show that evaluation depends on elicitation (Sclar et al., 2024; Polo et al., 2024). Pezeshkpour & Hruschka (2024) specifically establish option-order sensitivity in multiple-choice questions. Our order reversal pairs this known diagnostic with same-image attribute checks, complementing the contextual response analysis. The contribution is the linked comparison of context, observation, and elicitation, which exposes different behaviors behind similar marginal summaries.

3 ESCBench

3.1 Task and design principles

ESCBench separates changes to the *depicted situation* and its *observation* (Figure 1b) from changes to the *elicitation protocol* (Section 4). A simulated tabletop supports matched comparisons over shared scene geometry: we repeat each tool or layout contrast within each view, then compare its sign and distance profile across views. Each demonstrative task supplies one RGB image and identifies green-and-yellow MIMo as speaker and black-and-white iCub as addressee. Fixing one three-dimensional target, with tabletop markers treated as background, limits variation in target search. T1 restricts generation to *this* or *that*; T2 adds color and shape to measure recognition in the same response. Both request the speaker's perspective regardless of the camera. Coordinates, filenames, scene labels, and human probabilities are withheld so that responses use image and language evidence rather than privileged annotations.

3.2 HUMAN-INFORMED SCENE AND OBSERVATION DESIGN

Human studies motivate distance, tool, and interlocutor contrasts as complementary tests of situated reference (Figure 1a–b; Coventry et al., 2008; 2023). We adapt these contrasts, not participants' action experience: humans physically pointed and named objects in a memory game, whereas models observe static depictions. MIMo is simulated in MuJoCo with externally supported bases (Mattern et al., 2023; Todorov et al., 2012).

Distance, tools, and layout. Twelve positions trace a response profile without imposing a binary near/far boundary. Target M_i lies $d_i = 150i$ mm from the near table edge, $i = 1, \ldots, 12$; this axial distance is not eye–target distance. The 150-mm spacing is a 0.6 body-scale mapping of the adult study's 250-mm spacing, not evidence of equivalent reaching ability. A distance-only policy could nevertheless ignore action and interlocutors. Bare-hand versus 520-mm held-rod configurations test whether the profile changes with depicted action possibilities; side-by-side versus opposite layouts test sensitivity to addressee placement beyond target distance. Crossing both factors permits these comparisons at matched positions and appearances, with roles fixed. Because tool states also change posture, contact, and occlusion, these are configuration comparisons, not pure causal estimates of action capability.

Appearance. Four colors support recognition checks and matched comparisons within saved geometry; three shapes broaden target geometry. Sphere diameter, cube side, and cylinder diameter/height are 70 mm and remain fixed across positions, avoiding a physical-size change with distance. Shape still changes geometry and potentially visibility; it is not a cosmetic intervention.

Observation. Left speaker-eye and addressee-eye cameras provide participant-centred observations; Exo and orthographic Top provide external views of the configuration. Querying each view separately tests observation dependence without multi-image fusion. The designated speaker remains fixed, but visible evidence need not: disagreement is not automatically a reference error. Exo/Top cameras match across tools; Ego gaze protocols differ. Prompt contrasts separately vary elicitation on identical images (Section 4).

3.3 MATCHED CONSTRUCTION AND EVIDENCE BOUNDARIES

Crossing tool, layout, position, and shape yields 144 geometry templates; four colors produce 576 scenes, and four views produce 2,304 formal images. Linked identities support traceable comparisons: colors share a template, views share a scene, and 24 position–layout families record shared construction ancestry. The package supplies factual attributes, geometry, roles, cameras, interaction states, quality flags, 288 tool-paired scene pairs, and 1,200 contrast groups. Rendering specifications and 1,152 right-eye audit frames are documented separately (Appendices G and I); audit frames are not evaluation inputs.

To check construction validity, source static checks cover target geometry, fingertip/rod-tip features, pointing direction, clearance, native constraints, and support for all templates. Bare-hand M01–M02 and tool M01–M05 satisfy contact tolerances; other states depict non-contact pointing. These checks establish saved construction states, not maximum reach, dynamics, or strict unreachability; historical witnesses retain their own domains (Appendix N).

Annotation scope likewise separates what is recorded from what can be scored. The 384 Top sphere/cylinder inputs are excluded from primary shape/joint scoring to avoid treating ambiguous projections as exact three-dimensional recognition tests; they remain in demonstrative/color evaluation. The 624 bounding-box discrepancy records are retained as annotation-extent findings, not proven image corruption (Appendix I). No current-image human demonstrative labels are supplied. Shared construction also precludes treating a random image split as independent generalization; the fixed 192-image panel was already used for development and diagnostics.

4 EVALUATION PROTOCOL

Evaluation goals. We assess four complementary dimensions (Figure 1c): *recognition* checks target attributes; *context response* measures distance, tool, and layout dependence; *choice sensitivity* measures changes across views, task forms, and option orders; and *human-reference resemblance*

compares conditional choice distributions with external English responses. We first check recognition and elicitation sensitivity, then estimate tool and layout contrasts separately by distance and view. Their signed profiles reveal whether response direction and location persist across observations and qualify external-reference scores.

Protocols and deployments. We analyze 22 deployments: 17 open-weight configurations and five API aliases. T1, T2, and the T1 order reversal (D1) cover all images. Four tasks share a 192-image diagnostic panel: attributes only (T0), T0 with role identification (D2), T2 context requesting only attributes (D3), and T2 order reversal (D4). Each deployment contributes $3 \times 2,304 + 4 \times 192 = 7,680$ items, totaling 168,960 responses: 165,630 strictly valid and 3,330 invalid. Retries and candidate scoring do not enlarge this denominator. Native preprocessing, precision, hardware, and reasoning settings differ; API identities are gateway-reported. Configuration matching and execution details are in Appendix A, with exact prompts in Appendix D.

Measurements. Strict parsing accepts only the declared response structure and vocabulary. Attribute accuracy uses all expected applicable items: 2,304 colors and 1,920 shapes/joint attributes for T2; 192 and 160 on the panel. Let $p(C)$ be the proportion choosing *this* among valid responses in condition C. Invalid outputs are not counted as *that*. For N jointly valid pairs, with n_{01} *that*-to-*this* and n_{10} reverse transitions,

$$\Delta = (n_{01} - n_{10})/N, \qquad F = (n_{01} + n_{10})/N. \tag{1}$$

Tool contrasts are held minus bare; layout contrasts are side-by-side minus opposite. Table 1 averages absolute T2 Exo contrasts over eight cells: four quadrants crossed with layout for Tool, or tool state for Layout. Distance is the negative mean within-layout slope over all twelve T2 bare-hand Exo positions (pp/m). View is the fraction of complete T2 scenes with any four-view disagreement; Task and Order use F for T1/T2 and T2/D4. Expected/valid denominators and signed decompositions are retained (Appendix F). These fixed-stimulus summaries are descriptive, without independent-image significance tests.

External references. We distinguish the *tool reference* (Coventry et al., 2008), covering two tools and four quadrants (Q1–Q4), from the *distance–layout reference* (Coventry et al., 2023), covering bare-hand responses in two layouts and three regions: R1={M01,M02}, R2={M06,M07}, and R3={M11,M12}. JSD averages condition-wise square-root, base-2 Jensen–Shannon distances: eight tool–quadrant cells or six layout–region cells. Table 1 uses the latter in Exo. Human intervals resample participants. These are external distributions, not current-image labels. The human row takes Tool from the former study and Distance/Layout from the latter, on narrower supports than the model columns. Mappings and matched model recalculations are in Appendices E–F.

5 Results

5.1 RQ1: Recognition and choice stability

Table 1 separates recognition, context response, choice sensitivity, and human resemblance. Human and random-reference rows use the definitions in Section 4 and Appendix F.

Recognition and elicitation checks. The nine deployments with at least 99% joint attribute accuracy have T1–T2 choice-change rates of 4.7–75.2% (Figure 2a). Qwen3.8-27B scores perfectly on attributes yet changes 75.17% of choices; Mistral reaches 98.59% joint accuracy with 92.88% changes. These measure task-form sensitivity, since T1/T2 change both wording and response format.

Figure 2c distinguishes Mistral's task-dependent preference, Molmo2-ER's constant *this* responses, and Gemini's distance gradient with 5.64% task changes. Zero changes can therefore reflect insensitivity rather than contextual stability. These curves pool scene factors and views, unlike Table 1's bare-hand Exo slope.

View dependence survives recognition checks. Conditioning on correct scored attributes leaves four-view disagreement in 44.97% of Qwen3.8-27B scenes and 18.92% of Gemini scenes (Fig-

Table 1: **Evaluation results for 22 VLM deployments on ESCBench.** Bold marks eligible group-wise attribute maxima and JSD/sensitivity minima. Lower sensitivity means fewer changed choices, not necessarily better deixis. All runs are complete; choice statistics use parseable responses. Random guessing gives analytic expectations. Attributes and sensitivity are percentages; Distance uses pp/m and Tool/Layout use pp. Matched comparisons and metric details are reported separately.

Model	**Attributes**		**Context response**			**Choice sensitivity**			**Human ref.**
	Color ↑	Joint ↑	Distance	Tool	Layout	View ↓	Task ↓	Order ↓	JSD ↓
Proprietary (API)									
Claude Opus 5.5	97.2	96.4	75.7	5.6	1.4	42.5	8.4	22.4	0.323
GPT-5.6 Sol	99.8	99.7	72.6	2.1	15.3	28.6	5.8	6.3	0.292
GPT-6 Astra	98.4	99.1	75.7	3.8	13.5	33.9	10.3	10.9	**0.262**
GPT-6 Sol	**100.0**	99.9	79.3	1.7	12.8	33.7	**4.7**	**4.7**	0.295
Gemini 3.1 Pro Preview	**100.0**	**100.0**	71.2	2.4	14.9	**18.9**	5.6	5.7	0.304
Open-weight									
Qwen3-VL-2B	60.2	62.3	6.2	3.9	92.4	60.5	79.0	17.3	0.503
Qwen3.5-2B	66.8	68.2	0.0	0.0	0.0	0.0	100.0	100.0	0.693
InternVL-U	74.7	66.4	0.0	0.0	0.0	4.2	22.1	0.8	0.693
Gemma-3-4B	79.0	69.0	0.0	0.0	0.0	0.0	100.0	1.7	0.414
Qwen3.5-4B	99.4	99.3	73.5	8.3	11.1	48.4	36.5	12.5	0.376
Molmo2-ER	88.2	85.4	0.0	0.0	0.0	**0.0**	**0.0**	**0.5**	0.693
Gemma-4-E2B	95.1	86.0	60.4	20.8	27.8	71.9	48.9	56.8	**0.137**
Gemma-4-E4B	96.9	91.1	58.6	20.5	30.2	41.1	73.0	30.7	0.292
LLaVA-OV2-8B	96.9	96.3	8.8	7.6	18.1	58.0	32.2	21.4	0.512
InternVL3.5-8B	95.7	91.3	41.1	17.7	23.3	39.1	21.2	56.8	0.244
Cosmos3-Nano R.	99.7	99.7	56.6	14.9	5.9	62.0	38.8	10.4	0.410
Qwen3-VL-8B	97.7	97.3	0.0	0.0	0.0	7.1	2.5	8.3	0.414
Qwen3.8-27B	**100.0**	**100.0**	54.3	13.2	6.9	45.0	75.2	13.0	0.231
Gemma-4-31B	99.8	98.8	62.3	15.6	5.2	24.3	15.6	8.9	0.323
InternVL3.5-38B	**100.0**	99.7	44.4	8.3	10.4	25.5	35.5	58.3	0.286
Mistral-3.5-128B FP8	99.5	98.6	14.5	5.2	5.2	3.3	92.9	40.6	0.306
Qwen3.8-Flash FP8	**100.0**	99.9	73.7	7.6	8.3	33.9	35.5	11.5	0.257
Baselines									
Random guessing	25.0	8.3	0.0	9.4	9.4	87.5	50.0	50.0	0.276
Human reference	–	–	39.2	6.8	4.2	–	–	–	–

ure 2b). Recognition errors alone cannot explain this variation. Color-dependent choices also persist in 16.04% of Qwen's fully correct, shape-eligible groups (Appendix B.4). Views and colors can change evidence or salience, so these differences do not alone establish reference errors.

Option order changes choices on identical images. InternVL3.5-38B changes 112/192 T2–D4 choices (58.33%) while retaining perfect color accuracy. This comparison preserves the requested task, unlike T1/T2. Full-panel attribute and format transitions, other deployments, and configuration qualifications are reported in Appendix B.1; API paired differences can also include sampling variation.

5.2 RQ2: Tool and interlocutor responses depend on observation

Tool responses differ in sign and location. At M03–M05, where only the rod makes contact, InternVL3.5-38B's T2 held-minus-bare contrast is $+33.33$ pp in Exo and $+31.94$ pp in Top, whereas Qwen3.8-27B gives -16.67 pp in both views (Figure 3a). InternVL's response reverses at nearer positions and vanishes beyond M05. Gemini instead reverses across views at M04: $+12.50$ pp in Exo versus -33.33 pp in Top. An Exo-only reading of Gemini at M04 would associate the rod with

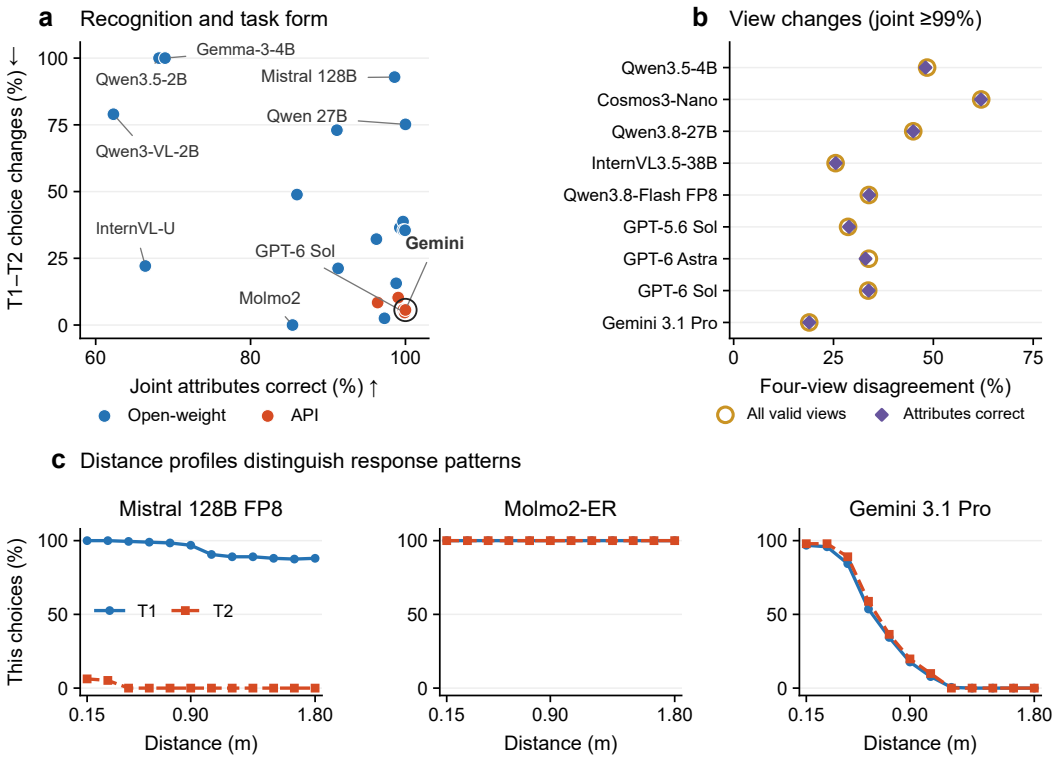


Figure 2: **Target recognition and choice stability are distinct.** (a) All 22 deployments: joint T2 accuracy versus T1/T2 changes on valid pairs. Arrows indicate higher accuracy or fewer changes, not overall quality. (b) All nine deployments with joint accuracy $\geq 99\%$: percentage of scenes with any disagreement among four valid T2 views. Gold circles use all complete scenes; purple diamonds require correct colors and eligible shapes in all views. (c) T1/T2 distance profiles pool scene factors and views (192 valid responses per point). This rate is not accuracy.

increased *this* use; adding Top changes that interpretation to an observation-dependent response of opposite sign. These are configuration associations, not isolated effects of action capability.

Layout responses depend on the observation. Figure 3b shows T2 layout contrasts for all five API deployments. At R2, GPT-5.6 Sol, Astra, and GPT-6 Sol range from -100 to -95.83 pp in addressee-eye views but give zero in speaker-eye views, despite MIMo's fixed speaker role. The other API deployments show the same concentration in the addressee-eye middle region, with smaller near/far responses. Speaker-eye alone suggests no R2 layout response for these three deployments; addressee-eye reveals a strong response under the same speaker. A separate T1 example illustrates aggregation: InternVL3.5-38B changes 47.66% of layout pairs, yet its net contrast is only $+0.43$ pp because opposing transitions cancel. Per-view counts and further contrasts are in Appendix B.4.

5.3 RQ3: External human resemblance is condition-specific

We compare conditional profiles using the separate tool and distance–layout references defined in Section 4.

A distance gradient is not sufficient for resemblance. The distance–layout reference has a graded near–far decline (Figure 4a). Gemini selects *this* for every R1 Exo input and none in R3, with at most 4.17% in R2; despite the shared near–far direction, its JSD is 0.304 (b). Qwen3.8-27B is closer overall (0.231), but its R2 choices differ markedly by layout: 37.50% opposite versus 4.17%

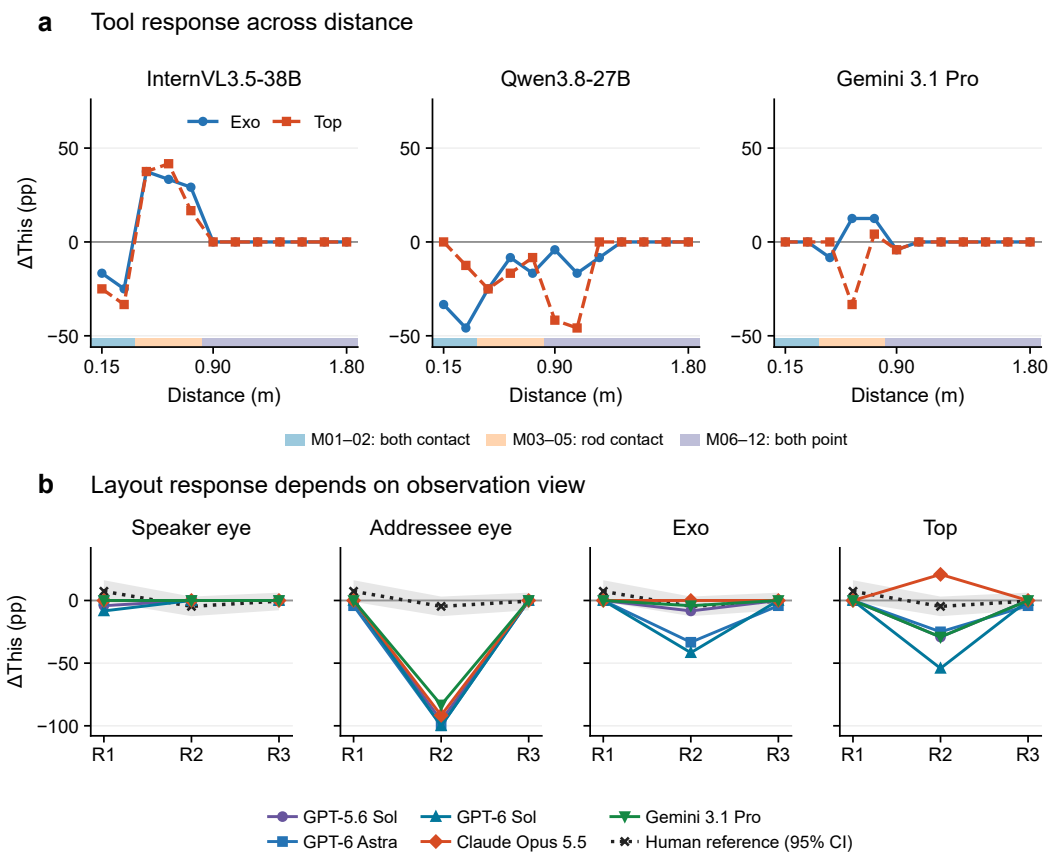


Figure 3: **Regional responses depend on condition and observation.** (a) T2 held-minus-bare contrasts in Exo/Top: 24 pairs per distance, balanced across layouts, shapes and colors. Strips show final static states, not reachability regions. (b) T2 bare-hand, side-by-side-minus-opposite contrasts for all five API deployments, with MIMo as speaker (24 pairs per point). The repeated human mean and 95% participant-bootstrap band are external English layout contrasts (Coventry et al., 2023), not responses to four rendered views.

side-by-side (c). It also has greater four-view disagreement than Gemini (44.97% versus 18.92%). A lower aggregate distance therefore need not imply greater agreement on each contextual contrast or greater view stability.

Similar scores can hide different contextual responses. Against the tool reference, Qwen3.8 Flash FP8 and GPT-6 Sol both have pooled-view JSD approximately 0.1405, yet their Q2 tool contrasts are $+22.57$ and -1.74 pp, respectively, versus the human $+22.22$ pp (Figure 4d). JSD alone would place the deployments equally close to the reference; the Q2 estimates distinguish a large positive shift, close to the human contrast, from a near-zero shift. This local comparison does not establish an overall ranking.

Pooling can create a closer match than any individual view. Qwen3.8-27B's distance–layout JSD falls to 0.133 after pooling, below all individual-view values (0.231–0.400; Figure 4e). Absolute-error comparisons show the same pattern (Appendix B.4). The pooled score describes a mixture unmatched by any constituent view, distinguishing resemblance within an observation from resemblance obtained by combining observations.

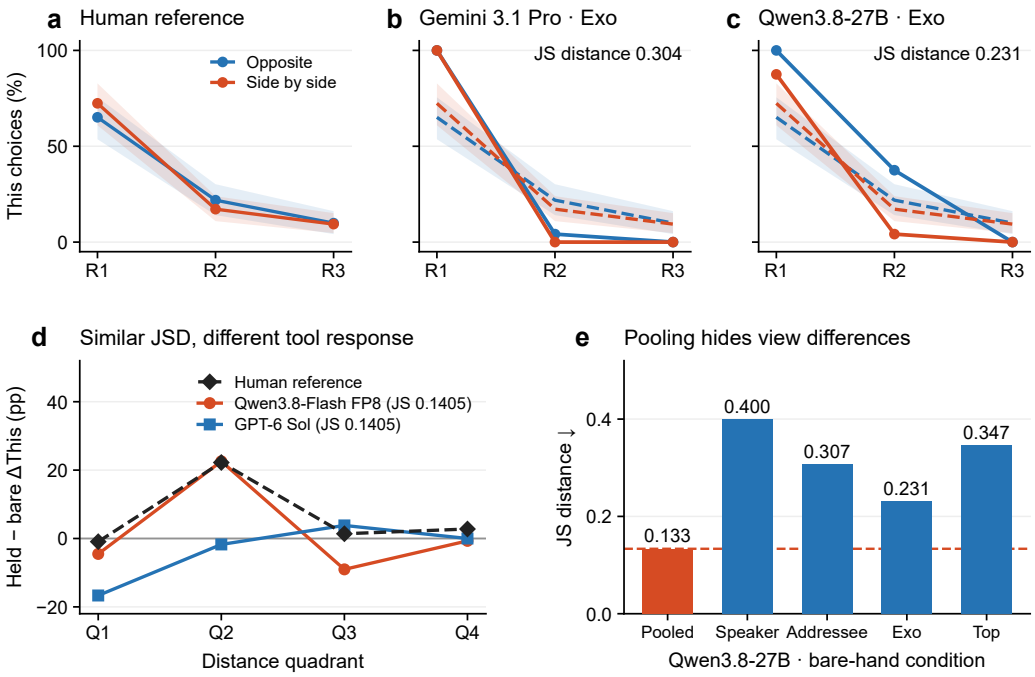


Figure 4: **Human resemblance depends on the comparison.** (a) English means and participant-bootstrap 95% intervals from Coventry et al. (2023). (b,c) T2/bare/Exo model profiles (solid) over human means (dashed) and intervals (shaded); 24 responses per cell. (d) Published human tool contrasts (Coventry et al., 2008) and model contrasts pooled over views/layouts (288 model pairs per Q). (e) Qwen3.8-27B's JSD to the later reference, per view (144 responses) and pooled (576). All displayed model responses are valid; the study-specific JSD definitions remain separate.

6 Conclusion

ESCBench repeats tool and layout contrasts within each view to compare response direction and distance profiles across observations under a fixed speaker task. Across 22 deployments, this separates behaviors that marginal summaries leave unresolved: a tool response can reverse across views despite accurate overall recognition and modest overall task-form sensitivity; similar human-reference scores can accompany a large positive response or a near-zero one; and pooling can yield a closer match than any single view. Constant outputs provide a further explanation for apparent stability.

The benchmark is limited to static, English-only tabletop reference, not enacted interaction. Tools covary with posture, contact, and occlusion; viewpoints change visible evidence; and external human distributions are not image-level answers. Choice changes alone neither establish reference errors nor identify internal mechanisms; visual shortcuts remain possible. Heterogeneous deployments and reuse of the diagnostic panel limit interpretation.

Same-image human judgments would calibrate interpretation, while repeated, configuration-matched evaluations would help separate elicitation sensitivity from sampling variation. Independent scenes, additional languages, and interactive tests would examine transfer. These extensions can test whether contextual responses persist across observations beyond the current controlled setting.

AI USE STATEMENT

Generative AI assisted research planning, simulation/control and evaluation code, data packaging, literature discovery, manuscript drafting and revision, statistical analysis scripts, and figure composition. The analysis uses recorded experimental outputs and scripted quantitative plots; generative image synthesis was not used to alter scientific results. Authors remain responsible for the validity, attribution, and final verification of the work.

REPRODUCIBILITY STATEMENT

The accompanying materials document the user prompts, deployment settings, derived results, and analysis procedures. We reparse all 168,960 recorded responses, verify stimulus correspondence and fixed-panel membership, and recompute paired and external-human summaries. This analysis does not repeat model inference. Earlier physical validation is reported with its original scope. Model weights, raw input tensors, and a complete rendering environment are not included in the manuscript package. Low-level execution identifiers and file-level provenance are retained in the accompanying technical records rather than reproduced in the paper.

ETHICS AND AVAILABILITY

The stimuli are synthetic; no new human-participant data were collected. External human data are analyzed in aggregate without treating them as image-level labels.

We will publicly release ESCBench upon acceptance, including all 2,304 benchmark images, scene and image annotations, evaluation prompts, evaluation code, and documentation. An anonymized version will be provided for peer review. The public release will be freely accessible and include version identifiers and citation information.

REFERENCES

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, et al. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a3cf318fbeec1126da21e9185ae9908c-Abstract-Conference.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, 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*, 2024. URL https://spatial-vlm.github.io/.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Kenny R. Coventry, Berenice Valdés, Alejandro Castillo, and Pedro Guijarro-Fuentes. Language within your reach: Near–far perceptual space and spatial demonstratives. *Cognition*, 108(3):889–895, 2008. doi: 10.1016/j.cognition.2008.06.010. URL https://www.kennycoventry.org/pdfs/cognition2008pub.pdf.

Kenny R. Coventry, Harmen B. Gudde, Holger Diessel, et al. Spatial communication systems across languages reflect universal action constraints. *Nature Human Behaviour*, 7(12):2099–2110, 2023. doi: 10.1038/s41562-023-01697-4. URL https://www.nature.com/articles/s41562-023-01697-4.

Dota Tianai Dong, Yifan Luo, Po-Ya Angela Wang, Asli Özyürek, and Paula Rubio-Fernández. Using perspectival words is harder than vocabulary words for humans—and even more so for multimodal language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1845–1870. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.84. URL https://aclanthology.org/2026.acl-long.84/.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T. Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

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*, 2023. URL https://aclanthology.org/2023.emnlp-main.568/.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32803–32827. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1514. URL https://aclanthology.org/2026.acl-long.1514/.

Ziqiao Ma, Jing Ding, Xuejun Zhang, Dezhi Luo, Jiahe Ding, Sihan Xu, Yuchen Huang, Run Peng, and Joyce Chai. Vision-language models are not pragmatically competent in referring expression

generation. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.16060v3.

Dominik Mattern, Pierre Schumacher, Francisco M. López, Marcel C. Raabe, Markus R. Ernst, Arthur Aubret, and Jochen Triesch. MIMo: A multi-modal infant model for studying cognitive development. arXiv:2312.04318, 2023. URL https://arxiv.org/abs/2312.04318.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28236482f64a72eec43706b6f3a6c511-Abstract-Conference.html.

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://research.nvidia.com/publication/2025-06_robospatial-teaching-spatial-understanding-2d-and-3d-vision-language-models.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL*, 2024. URL https://aclanthology.org/2024.findings-acl.719/.

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*, 2022. URL https://arxiv.org/abs/2204.03162.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109. URL https://roboti.us/lab/papers/TodorovIROS12.pdf.

Kaito Watanabe, Taisei Yamamoto, Tomoki Doi, and Hitomi Yanaka. Evaluation of multilingual ability to use spatial deictic expressions in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1203–1211. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-srw.106. URL https://aclanthology.org/2026.acl-srw.106/.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fe5d4bd3e2af823701b5d8f70a3c7602-Abstract-Conference.html.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025a. URL https://aclanthology.org/2025.acl-long.568/.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/af2d9fb5bcee19ef2dfa70d843520c97-Abstract-Conference.html.

Appendix Contents

A Model configurations and evaluation coverage **16**

A.1 Deployment identity and comparison scope 16

A.2 Mistral inference configuration and response checks 17

A.3 Response validation and recovery 17

B Complete deployment results **18**

B.1 Fixed 192-image diagnostic panel 18

B.2 Likelihood versus generated frequency 18

B.3 Analytic reference policies 23

B.4 Additional conditional diagnostics 25

C Distance profiles for all deployments **25**

D Evaluation prompts **29**

D.1 T0: attributes only 29

D.2 T1: demonstrative only 29

D.3 T2: referring phrase 29

D.4 D1: reversed demonstrative order in T1 29

D.5 D2: attribute recognition with participant roles 29

D.6 D3: attributes with full scene context 30

D.7 D4: reversed demonstrative order in T2 30

E External human comparison: design and estimands **30**

E.1 Coventry 2008: published means 30

E.2 Coventry 2023: English trial data 30

E.3 Distribution distance and signed contrasts 31

F Summary metrics and complete decompositions **31**

G Appearance specifications and reference renders **36**

G.1 Appearance family and reference renders 36

G.2 Scale, color and rendering controls 36

G.3 Placement and historical projection estimates 37

G.4 Evidence scope after formal construction 38

H Formal construction: geometry, adaptation and layout **39**

H.1 Scene structure and geometric representation 39

H.2 Separate adaptation and acceptance rules 39

H.3 Coordinates, seating and retained cloth 39

I Formal views and annotation limits **40**

J Data organization and comparison units **41**

K Evaluation implementation **44**

K.1 Exact prompts and parsing 44

K.2 Metrics and denominators 44

K.3 External English reference records 44

L Reproducibility and benchmark availability **45**

M Transfer from human experiments and comparison with related benchmarks **45**

M.1 What is transferred 45

M.2 Relation to the closest evaluations 46

N Historical physical evidence: overview and exact scope **47**

N.1 Problem and evidence types 47

N.2 Dynamic and static positive witnesses 47

N.3 Conditional exclusion and expanded-domain unknowns for disks 48

N.4 Benchmark interpretation and scope 49

O M02 dynamic acceptance **50**

O.1 Actual contact protocol and measurement windows 50

O.2 Execution and supporting measurements 51

P Conditional M03 exclusion certificate **52**

P.1 Configuration domain and entire-surface enclosure 52

P.2 Coverage, final tolerance extension, and numerical audit 53

Q Static searches and retained limitations **54**

Q.1 Expanded-domain bare-hand M03 55

Q.2 M05 static positive and visual mismatch 55

Q.3 M06 with native posture ranges and a fixed grasp 57

R Protocol comparison and reuse **58**

S Historical static witnesses for the 70-mm blue-sphere task **60**

S.1 Target geometry and witness configurations 60

S.2 Tolerances, posture and grasp 60

S.3 Scope relative to the original evidence 61

APPENDIX GUIDE

The appendix first documents the evaluated model protocols, complete results, and exact user prompts. It then describes the current stimulus geometry, appearance, cameras, and comparison identities, followed by the transfer from human experiments. Historical disk dynamics, static searches, and conditional reach certificates appear last. Their target geometries, allowed postures, and evidence types differ from the formal image collection; they are not language labels for that collection.

This ordering distinguishes three sources of evidence: the formal stimuli used in model evaluation, the recorded model responses, and the historical physical studies supporting construction development. The historical technical sections retain their original quantitative details. Passing a static geometry check or finding a geometric witness is never counted as a correct demonstrative response.

A MODEL CONFIGURATIONS AND EVALUATION COVERAGE

Each of 22 deployments contributes 7,680 image–prompt evaluation items, totaling 168,960 model–task–image combinations. All items have responses: 165,630 strictly valid and 3,330 invalid. The stimuli comprise 2,304 images, 576 scenes, 144 geometry templates, and 24 dependency families. Evaluation items are neither independent scenes nor inference-call counts. We retain one response per model–task–image combination; technical retries replace failed responses rather than increasing the evaluation denominator.

Table 2: Registered coverage per deployment. All four panel tasks use exactly the same 192 image IDs; primary shape/joint scoring uses 160 of these.

Task	Items	Request / controlled wording change
T0	192	Color and shape only.
T1	2,304	Speaker-perspective choice, This or That.
T2	2,304	Demonstrative, color, shape.
D1	2,304	T1 candidate order reversed.
D2	192	T0 with role identification added.
D3	192	T2 role, target and viewpoint context; attributes only.
D4	192	T2 candidate order reversed.

A.1 DEPLOYMENT IDENTITY AND COMPARISON SCOPE

For each deployment, we document the evaluated model component, parameter count, numerical precision, image processing, and generation settings. Open-weight rows are sorted using recorded total parameters or the evaluated component inventory. Gemma E2B/E4B denote effective scales, not total storage; their recorded totals are approximately 5.12B/8.00B. InternVL-U's complete inventory and loaded understanding component are distinct. Cosmos3-Nano here is its evaluated reasoner component, not the complete multi-component system. Qwen3.8 Flash is MoE: total, active, and additional embedding/prediction components are separately recorded. API parameter counts remain unknown. This ordering does not support a controlled parameter-scaling claim.

The four early deployments (Qwen3-VL-2B/8B, Gemma-3-4B, InternVL3.5-8B) combine earlier main-task runs with later diagnostics. Their D1 and D4 comparisons include differences in hardware or execution settings; configuration matching is also documented for the panel's attribute comparisons. Matching model names or settings does not establish numerical equivalence across hardware. One earlier Qwen3-VL-2B raw generation record could not be recovered, although its exported response is available for analysis.

API deployments request high reasoning settings and 25,000-token output budgets through gateways, whose interfaces differ. Candidate-sequence scores are unavailable. Backend model identity and effective reasoning allocation are not independently verified from gateway labels. Repeated calls may also differ, so API option-order changes are paired observations rather than isolated causal effects of order. Technical retries can involve duplicate processing, but recovered responses replace earlier failures and do not enlarge the evaluation denominator.

Table 3: Deployment inventory. Valid/invalid counts use all 7,680 items. Scores count stored candidate evaluations, not additional images. Recorded B follows the full/component scope documented in the source registry; effective or active scales are separate.

Deployment	Recorded B	Precision	Valid	Invalid	Scores
Qwen3-VL-2B	2.13	BF16	7655	25	4608
Qwen3.5-2B	2.27	BF16	6913	767	4608
InternVL-U	4.28	BF16	6901	779	4608
Gemma-3-4B	4.30	BF16	7566	114	4608
Qwen3.5-4B	4.66	BF16	7680	0	4608
Molmo2-ER	4.85	BF16	7680	0	4608
Gemma-4-E2B	5.12	BF16	7668	12	4608
Gemma-4-E4B	8.00	BF16	7669	11	4608
LLaVA-OneVision-2-8B	8.53	BF16	6087	1593	4608
InternVL3.5-8B	8.53	BF16	7654	26	4608
Cosmos3-Nano Reasoner	8.77	BF16	7680	0	4608
Qwen3-VL-8B	8.77	BF16	7680	0	4608
Qwen3.8-27B	27.78	BF16	7680	0	4608
Gemma-4-31B	31.27	BF16	7680	0	4608
InternVL3.5-38B	38.39	BF16	7680	0	4608
Mistral-Medium-3.5-128B FP8	127.70	Mixed FP8/BF16	7678	2	4608
Qwen3.8-Flash-Next FP8	180.00	Mixed FP8/BF16	7680	0	4608
Claude Opus 5.5	unknown	API	7679	1	0
GPT-5.6 Sol	unknown	API	7680	0	0
GPT-6 Astra	unknown	API	7680	0	0
GPT-6 Sol	unknown	API	7680	0	0
Gemini 3.1 Pro Preview (API, high requested)	unknown	API	7680	0	0

A.2 Mistral inference configuration and response checks

We evaluate Mistral-Medium-3.5-128B using mixed FP8/BF16 weight precision, with a small FP32 component, and greedy decoding. Inference uses Transformers and PyTorch with scaled dot-product attention, batch size one, random seed zero, and a maximum of 32 generated tokens. Sampling, explicit thinking, speculative decoding, and multi-token prediction are disabled. The run uses two RTX PRO 6000 Blackwell Server Edition GPUs. Deterministic algorithms are not enforced, so greedy decoding does not guarantee bitwise-identical results.

The native image processor resizes the longest edge to 1,540 pixels and uses 14-pixel patches, producing 868×1540 RGB inputs for the benchmark images. Image budgets and numerical precision are not matched across models. The model's nominal size and measured parameter count are reported separately in the deployment inventory.

The Mistral records contain 7,680 benchmark generations and 4,608 candidate-score evaluations. Text-only controls and numerical checks are excluded from benchmark item counts. Stored generations are matched to the corresponding tasks, images, and configurations. Recomputing candidate probabilities and log-odds from the recorded branch log probabilities reproduces all 4,608 score entries; summing token log probabilities also reproduces all 9,216 branch totals. These checks assess stored outputs rather than repeat model inference.

Recorded preprocessing information agrees for all 2,304 T1/T2 pairs, 2,304 T1/D1 pairs, and 192 T2/D4 pairs. The original pixel tensors and fully expanded runtime chat messages were not reloaded for this analysis, so input-matching evidence relies on the saved execution records. The seven user prompts are provided in Appendix D.

A.3 Response validation and recovery

We reparse all 168,960 stored responses, verify their correspondence to stimulus metadata, check uniqueness, and confirm all 88 deployment–panel-task memberships. We then recompute expected and valid denominators, paired task and condition contrasts, attribute and format transitions, and the two external human comparisons. This analysis uses recorded outputs rather than repeating inference or simulation.

Evaluation corpus. The analysis uses the response collection completed on 25 September 2026. Gemini contributes 7,680 strictly valid answers. Its initial 7,522 responses include 12 reused pilot responses; technical retries supply the remaining 158. Repeated requests are not independent eval-

Table 4: Exact valid-pair denominators and directional choice changes. T1 has 2,304 expected items; T1/T2 has 2,304 expected pairs; D4 has 192. Letters T/t in the final columns denote This/That.

Deployment	T1 valid	T1/T2 pairs	This→That	That→This	D4 T→t	D4 t→T
Qwen3-VL-2B	2304	2292	0	1810	0	31
Qwen3.5-2B	2304	1595	0	1595	127	0
InternVL-U	2304	1863	14	398	0	1
Gemma-3-4B	2304	2208	2208	0	0	3
Qwen3.5-4B	2304	2304	840	0	1	23
Molmo2-ER	2304	2304	0	0	1	0
Gemma-4-E2B	2304	2304	132	994	109	0
Gemma-4-E4B	2304	2304	1682	0	59	0
LLaVA-OneVision-2-8B	2290	2290	737	0	0	41
InternVL3.5-8B	2304	2304	89	400	0	109
Cosmos3-Nano Reasoner	2304	2304	879	14	1	19
Qwen3-VL-8B	2304	2304	18	40	0	16
Qwen3.8-27B	2304	2304	1732	0	2	23
Gemma-4-31B	2304	2304	324	36	0	17
InternVL3.5-38B	2304	2304	784	35	0	112
Mistral Medium 3.5 FP8	2304	2304	2140	0	0	78
Qwen3.8 Flash FP8	2304	2304	773	45	0	22
Claude Opus 5.5	2304	2304	83	111	41	2
GPT-5.6 Sol	2304	2304	83	50	5	7
GPT-6 Astra	2304	2304	111	126	15	6
GPT-6 Sol	2304	2304	82	26	7	2
Gemini 3.1 Pro Preview	2304	2304	47	83	7	4

uation items. The gateway reports the model alias but does not independently establish the backend model or effective reasoning allocation; candidate-sequence scores are unavailable.

B Complete deployment results

The accompanying machine-readable tables contain every task's expected, present, valid, invalid, and no-answer counts; applicable shape counts; separate color/shape/joint numerators; and every condition's valid denominator. No invalid response is interpreted as That. Low valid coverage can make a zero-flip estimate uninformative: LLaVA-OneVision-2-8B's D1 has only 725 valid outputs out of 2,304 expected items. All invalid outputs remain recorded.

B.1 Fixed 192-image diagnostic panel

The panel is the existing development/diagnostic subset, not a new held-out set. Every comparison below pairs responses to the same images. Four-state attribute transitions preserve all 192 expected color items and 160 primary shape/joint items. Missing/invalid outputs count as not correct in these expected-denominator tables; separate both-valid analyses, shape-only transitions, and full valid/invalid/no-answer state matrices accompany them. Equal aggregate accuracy need not mean the same individual images remain correct. All registered items have effective answers; invalid transitions in other deployments remain substantive. D1/D4 comparisons use T1/D1 restricted to these same 192 images, alongside T2/D4, rather than subtracting the full-pool D1 rate from a panel D4 rate. Configuration matching is checked separately for every comparison. Format and attribute changes under D4 are retained even when a demonstrative is parseable; the public metric requires the full declared output structure to be valid.

B.2 Likelihood versus generated frequency

Open-weight candidate evaluations compare complete This and That continuations with the native assistant ending, summed token log probabilities, and two-sequence normalization, without length averaging. Each score is associated with its prompt and candidate-extraction procedure. These conditional scores are neither T2 accuracy nor human response probabilities. For example, the historical Qwen3-VL-2B always generates That in T1 while its stored candidate scores vary with distance. Generation saturation therefore need not imply identical candidate preferences.

Scores from different prompts or model-specific extraction procedures are analyzed separately. Candidate statistics are separate from the generated-choice analysis and are not used for the external

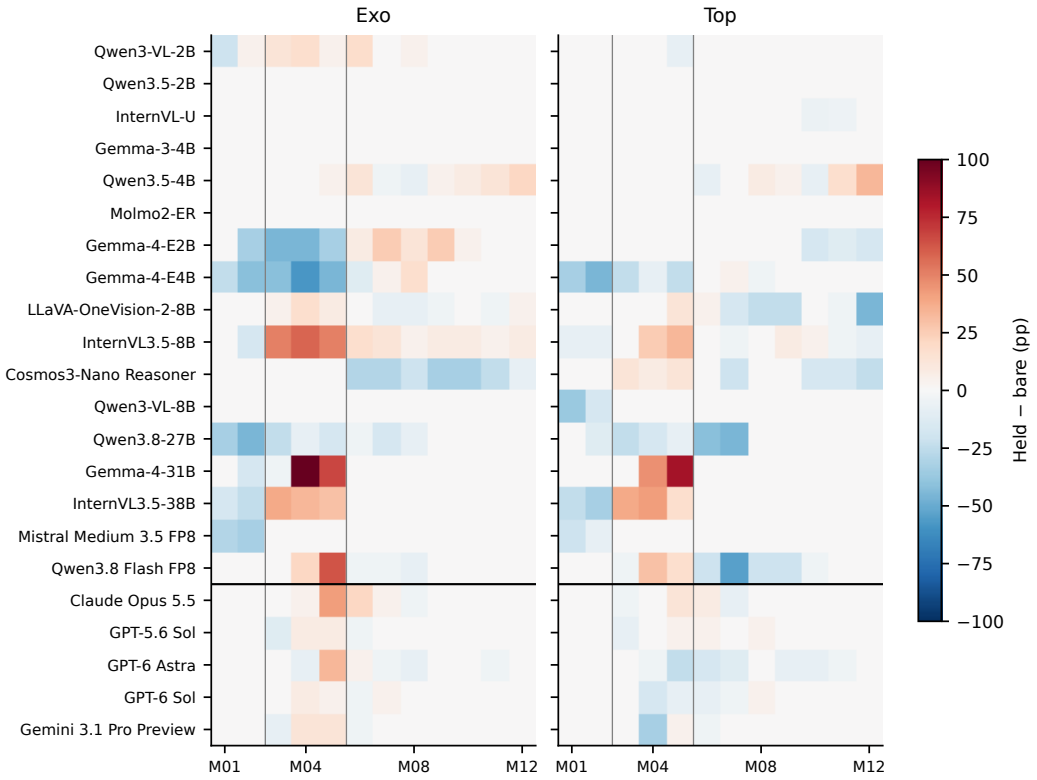


Figure 5: T2 held-minus-bare tool contrasts for all twelve positions and all deployments, separately for the camera-matched Exo and Top views. There are 24 expected pairs per cell. Vertical dividers locate the final static-state strata; they do not certify reach boundaries. Exact valid coverage is retained in the position-wise table.

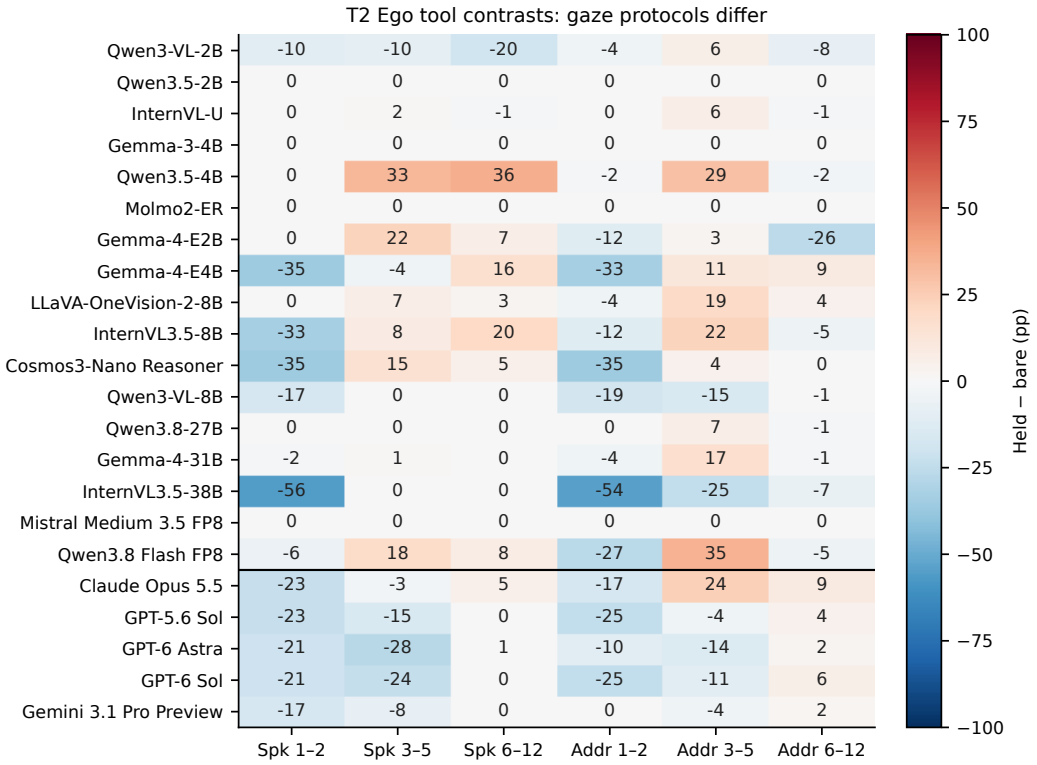


Figure 6: All-deployment T2 Ego tool contrasts, separately by view and final-state stratum. These camera protocols differ between bare and held conditions; they are not camera-matched effects. Positions 1–2, 3–5, and 6–12 have 48, 72, and 168 expected pairs per view respectively. Exact valid denominators remain in the source table.

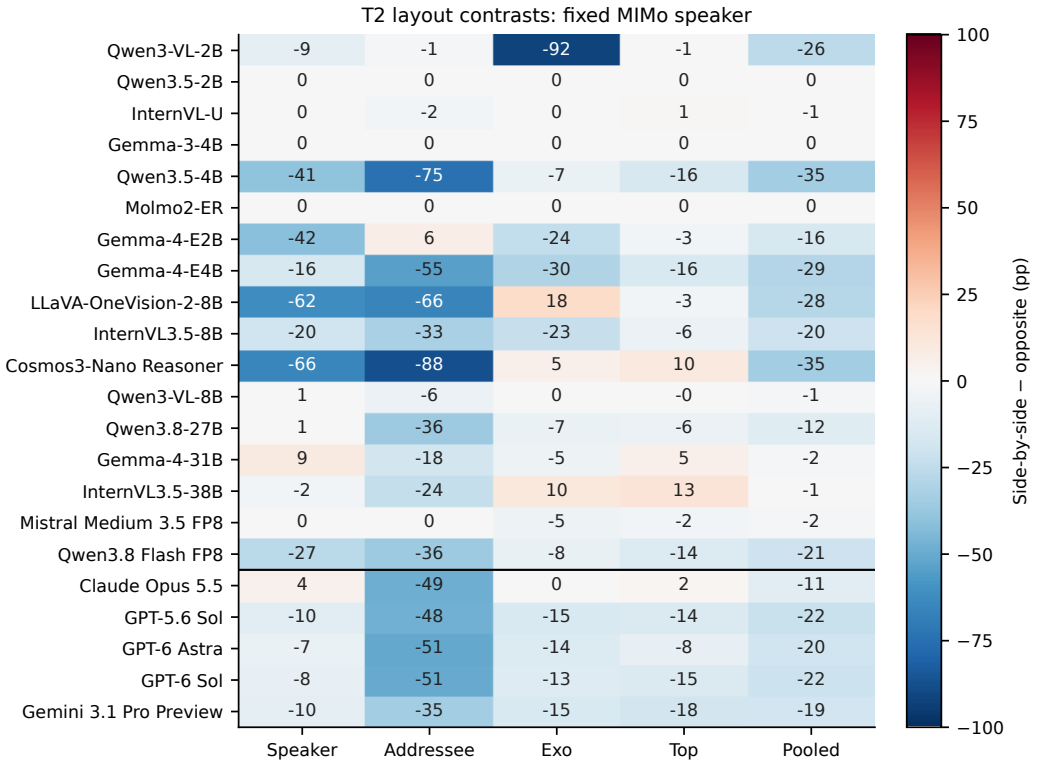


Figure 7: All-deployment T2 layout contrasts, side-by-side minus opposite. Each view has 288 expected pairs; the pool has 1,152. Values use jointly valid pairs, with coverage and bidirectional switching reported separately in the accompanying tables.

Table 5: T2 external-reference summaries, with views pooled. JS is macro base-2 Jensen–Shannon distance; MAE is in percentage points. The 2008 comparison covers eight tool/quadrant cells; 2023 covers six bare-hand layout/region cells. Their reference sources, mappings, and conditions differ, so the columns are not pooled into one score. Condition-level and separate-view results accompany the source.

Deployment	JS 08	MAE 08	Valid 08	JS 23	MAE 23	Valid 23
Qwen3-VL-2B	0.474	51.49	2292/2304	0.484	50.21	573/576
Qwen3.5-2B	0.740	72.51	1595/2304	0.693	67.36	395/576
InternVL-U	0.714	71.46	1863/2304	0.689	67.14	476/576
Gemma-3-4B	0.360	27.49	2208/2304	0.414	32.64	556/576
Qwen3.5-4B	0.372	36.05	2304/2304	0.321	28.99	576/576
Molmo2-ER	0.740	72.51	2304/2304	0.693	67.36	576/576
Gemma-4-E2B	0.398	41.26	2304/2304	0.363	37.85	576/576
Gemma-4-E4B	0.106	10.30	2304/2304	0.208	14.93	576/576
LLaVA-OneVision-2-8B	0.408	40.26	2304/2304	0.380	36.28	576/576
InternVL3.5-8B	0.093	7.78	2304/2304	0.144	11.46	576/576
Cosmos3-Nano Reasoner	0.327	33.14	2304/2304	0.387	36.63	576/576
Qwen3-VL-8B	0.321	25.15	2304/2304	0.321	26.22	576/576
Qwen3.8-27B	0.115	8.55	2304/2304	0.133	9.72	576/576
Gemma-4-31B	0.158	11.69	2304/2304	0.282	18.75	576/576
InternVL3.5-38B	0.178	12.99	2304/2304	0.172	11.28	576/576
Mistral Medium 3.5 FP8	0.343	26.53	2304/2304	0.364	28.82	576/576
Qwen3.8 Flash FP8	0.140	13.40	2304/2304	0.216	16.15	576/576
Claude Opus 5.5	0.138	12.10	2304/2304	0.241	15.80	576/576
GPT-5.6 Sol	0.134	10.46	2304/2304	0.277	18.23	576/576
GPT-6 Astra	0.125	11.29	2304/2304	0.261	18.92	576/576
GPT-6 Sol	0.141	11.11	2304/2304	0.295	20.66	576/576
Gemini 3.1 Pro Preview	0.172	12.98	2304/2304	0.277	17.71	576/576

Table 6: T0 / D2: paired attribute outcomes, in order both correct / A only / B only / neither. All expected items remain in denominators. † denotes mixed deployment configurations; ‡ denotes API comparisons. Full both-valid and format transitions are supplied in the accompanying data tables.

Deployment	Color (192)	Joint (160)	Invalid A/B
Qwen3-VL-2B†	183/8/0/1	151 / 7 / 2 / 0	0/0
Qwen3.5-2B	188/3/0/1	156 / 3 / 0 / 1	1/4
InternVL-U	41/4/38/109	38 / 4 / 31 / 87	147/100
Gemma-3-4B	183/2/1/6	112 / 6 / 4 / 38	3/2
Qwen3.5-4B	192/0/0/0	160 / 0 / 0 / 0	0/0
Molmo2-ER	192/0/0/0	158 / 0 / 0 / 2	0/0
Gemma-4-E2B	184/2/0/6	138 / 3 / 1 / 18	5/7
Gemma-4-E4B	181/1/4/6	142 / 0 / 3 / 15	5/3
LLaVA-OneVision-2-8B	192/0/0/0	160 / 0 / 0 / 0	0/0
InternVL3.5-8B	160/1/19/12	135 / 0 / 13 / 12	17/9
Cosmos3-Nano Reasoner	192/0/0/0	160 / 0 / 0 / 0	0/0
Qwen3-VL-8B	192/0/0/0	160 / 0 / 0 / 0	0/0
Qwen3.8-27B	192/0/0/0	160 / 0 / 0 / 0	0/0
Gemma-4-31B	192/0/0/0	156 / 0 / 1 / 3	0/0
InternVL3.5-38B	192/0/0/0	159 / 0 / 0 / 1	0/0
Mistral Medium 3.5 FP8	191/0/0/1	154 / 1 / 1 / 4	1/1
Qwen3.8 Flash FP8	192/0/0/0	160 / 0 / 0 / 0	0/0
Claude Opus 5.5‡	188/2/0/2	155 / 2 / 0 / 3	0/0
GPT-5.6 Sol‡	192/0/0/0	159 / 0 / 1 / 0	0/0
GPT-6 Astra‡	191/0/1/0	159 / 0 / 1 / 0	0/0
GPT-6 Sol‡	192/0/0/0	159 / 1 / 0 / 0	0/0
Gemini 3.1 Pro Preview‡	192/0/0/0	160 / 0 / 0 / 0	0/0

human scores in this study. APIs lack a verified corresponding interface. Mistral score checks use stored token and branch probabilities rather than repeated forward evaluation.

Table 7: T2 / D3: paired attribute outcomes, in order both correct / A only / B only / neither. All expected items remain in denominators. † denotes mixed deployment configurations; ‡ denotes API comparisons. Full both-valid and format transitions are supplied in the accompanying data tables.

Deployment	Color (192)	Joint (160)	Invalid A/B
Qwen3-VL-2B†	123/1/33/35	106 / 2 / 24 / 28	1/0
Qwen3.5-2B	106/17/17/52	83 / 17 / 15 / 45	65/35
InternVL-U	101/42/16/33	77 / 29 / 12 / 42	41/38
Gemma-3-4B	129/23/17/23	82 / 25 / 17 / 36	9/0
Qwen3.5-4B	189/0/2/1	157 / 0 / 2 / 1	0/0
Molmo2-ER	149/19/2/22	120 / 14 / 2 / 24	0/0
Gemma-4-E2B	171/5/8/8	125 / 4 / 11 / 20	0/0
Gemma-4-E4B	181/3/0/8	137 / 3 / 4 / 16	0/3
LLaVA-OneVision-2-8B	182/2/2/6	150 / 2 / 2 / 6	0/0
InternVL3.5-8B	179/2/1/10	140 / 2 / 2 / 16	0/0
Cosmos3-Nano Reasoner	190/1/1/0	158 / 1 / 1 / 0	0/0
Qwen3-VL-8B	183/0/6/3	151 / 0 / 6 / 3	0/0
Qwen3.8-27B	192/0/0/0	160 / 0 / 0 / 0	0/0
Gemma-4-31B	192/0/0/0	155 / 1 / 1 / 3	0/0
InternVL3.5-38B	192/0/0/0	158 / 0 / 1 / 1	0/0
Mistral Medium 3.5 FP8	190/0/2/0	153 / 2 / 3 / 2	0/0
Qwen3.8 Flash FP8	192/0/0/0	160 / 0 / 0 / 0	0/0
Claude Opus 5.5‡	186/1/1/4	153 / 1 / 1 / 5	0/0
GPT-5.6 Sol‡	189/3/0/0	158 / 0 / 1 / 1	0/0
GPT-6 Astra‡	182/7/2/1	153 / 5 / 2 / 0	0/0
GPT-6 Sol‡	191/1/0/0	158 / 1 / 0 / 1	0/0
Gemini 3.1 Pro Preview‡	192/0/0/0	160 / 0 / 0 / 0	0/0

Table 8: T0 / T2: paired attribute outcomes, in order both correct / A only / B only / neither. All expected items remain in denominators. † denotes mixed deployment configurations; ‡ denotes API comparisons. Full both-valid and format transitions are supplied in the accompanying data tables.

Deployment	Color (192)	Joint (160)	Invalid A/B
Qwen3-VL-2B	124/67/0/1	106 / 52 / 2 / 0	0/1
Qwen3.5-2B	123/68/0/1	100 / 59 / 0 / 1	1/65
InternVL-U	41/4/102/45	36 / 6 / 70 / 48	147/41
Gemma-3-4B	151/34/1/6	96 / 22 / 11 / 31	3/9
Qwen3.5-4B	189/3/0/0	157 / 3 / 0 / 0	0/0
Molmo2-ER	168/24/0/0	134 / 24 / 0 / 2	0/0
Gemma-4-E2B	173/13/3/3	125 / 16 / 4 / 15	5/0
Gemma-4-E4B	179/3/5/5	137 / 5 / 3 / 15	5/0
LLaVA-OneVision-2-8B	184/8/0/0	152 / 8 / 0 / 0	0/0
InternVL3.5-8B	150/11/31/0	129 / 6 / 13 / 12	17/0
Cosmos3-Nano Reasoner	191/1/0/0	159 / 1 / 0 / 0	0/0
Qwen3-VL-8B	183/9/0/0	151 / 9 / 0 / 0	0/0
Qwen3.8-27B	192/0/0/0	160 / 0 / 0 / 0	0/0
Gemma-4-31B	192/0/0/0	156 / 0 / 0 / 4	0/0
InternVL3.5-38B	192/0/0/0	158 / 1 / 0 / 1	0/0
Mistral Medium 3.5 FP8	189/2/1/0	153 / 2 / 2 / 3	1/0
Qwen3.8 Flash FP8	192/0/0/0	160 / 0 / 0 / 0	0/0
Claude Opus 5.5‡	187/3/0/2	154 / 3 / 0 / 3	0/0
GPT-5.6 Sol‡	192/0/0/0	158 / 1 / 0 / 1	0/0
GPT-6 Astra‡	188/3/1/0	157 / 2 / 1 / 0	0/0
GPT-6 Sol‡	192/0/0/0	159 / 1 / 0 / 0	0/0
Gemini 3.1 Pro Preview‡	192/0/0/0	160 / 0 / 0 / 0	0/0

B.3 Analytic reference policies

For $y_\tau(k) = \mathbf{1}[k \le \tau]$, independent of prompt, tool, layout, color and view, constants $\tau = 0, 12$ have zero switching and zero distance slope and zero near–far contrast. For $2 \le \tau \le 10$, the rule has a maximal endpoint near–far contrast and perfect agreement but zero tool/layout sensitivity. These are transparent analytic examples, not experimental VLM competitors or human labels. They explain why no single recognition, gradient, or consistency statistic establishes situated reference.

Table 9: T2 / D4: paired attribute outcomes, in order both correct / A only / B only / neither. All expected items remain in denominators. † denotes mixed deployment configurations; ‡ denotes API comparisons. Full both-valid and format transitions are supplied in the accompanying data tables.

Deployment	Color (192)	Joint (160)	Invalid A/B
Qwen3-VL-2B†	123/1/13/55	107 / 1 / 9 / 43	1/13
Qwen3.5-2B	123/0/36/33	100 / 0 / 34 / 26	65/18
InternVL-U	122/21/10/39	90 / 16 / 11 / 43	41/53
Gemma-3-4B†	143/9/1/39	102 / 5 / 3 / 50	9/13
Qwen3.5-4B	189/0/1/2	157 / 0 / 1 / 2	0/0
Molmo2-ER	168/0/0/24	134 / 0 / 0 / 26	0/0
Gemma-4-E2B	173/3/5/11	127 / 2 / 6 / 25	0/0
Gemma-4-E4B	181/3/1/7	138 / 2 / 1 / 19	0/0
LLaVA-OneVision-2-8B	180/4/1/7	148 / 4 / 1 / 7	0/0
InternVL3.5-8B†	177/4/1/10	138 / 4 / 2 / 16	0/0
Cosmos3-Nano Reasoner	191/0/0/1	159 / 0 / 0 / 1	0/0
Qwen3-VL-8B†	183/0/2/7	151 / 0 / 2 / 7	0/0
Qwen3.8-27B	192/0/0/0	160 / 0 / 0 / 0	0/0
Gemma-4-31B	192/0/0/0	156 / 0 / 0 / 4	0/0
InternVL3.5-38B	192/0/0/0	158 / 0 / 0 / 2	0/0
Mistral Medium 3.5 FP8	190/0/1/1	155 / 0 / 1 / 4	0/0
Qwen3.8 Flash FP8	192/0/0/0	160 / 0 / 0 / 0	0/0
Claude Opus 5.5‡	187/0/4/1	150 / 4 / 5 / 1	0/0
GPT-5.6 Sol‡	192/0/0/0	158 / 0 / 0 / 2	0/0
GPT-6 Astra‡	185/4/3/0	156 / 2 / 2 / 0	0/0
GPT-6 Sol‡	192/0/0/0	159 / 0 / 0 / 1	0/0
Gemini 3.1 Pro Preview‡	192/0/0/0	160 / 0 / 0 / 0	0/0

Table 10: Order diagnostics on the common 192-image panel. Directional columns are This-to-That / That-to-This counts. Config describes T2/D4; configuration details for each comparison are documented separately.

Deployment	D1 valid	D1 T–t/t–T	D4 valid	D4 T–t/t–T	Config
Qwen3-VL-2B	192/192	0/0	179/192	0/31	mixed
Qwen3.5-2B	192/192	0/192	127/192	127/0	matched
InternVL-U	192/192	24/31	130/192	0/1	matched
Gemma-3-4B	192/192	181/0	179/192	0/3	mixed
Qwen3.5-4B	192/192	0/0	192/192	1/23	matched
Molmo2-ER	192/192	0/0	192/192	1/0	matched
Gemma-4-E2B	192/192	0/117	192/192	109/0	matched
Gemma-4-E4B	192/192	8/0	192/192	59/0	matched
LLaVA-OneVision-2-8B	61/192	0/0	192/192	0/41	matched
InternVL3.5-8B	192/192	0/160	192/192	0/109	mixed
Cosmos3-Nano Reasoner	192/192	0/1	192/192	1/19	matched
Qwen3-VL-8B	192/192	0/188	192/192	0/16	mixed
Qwen3.8-27B	192/192	65/1	192/192	2/23	matched
Gemma-4-31B	192/192	54/0	192/192	0/17	matched
InternVL3.5-38B	192/192	18/34	192/192	0/112	matched
Mistral-Medium-3.5-128B FP8	192/192	84/3	192/192	0/78	matched
Qwen3.8-Flash-Next FP8	192/192	0/41	192/192	0/22	matched
Claude Opus 5.5	192/192	12/3	192/192	41/2	API
GPT-5.6 Sol	192/192	3/6	192/192	5/7	API
GPT-6 Astra	192/192	12/13	192/192	15/6	API
GPT-6 Sol	192/192	4/3	192/192	7/2	API
Gemini 3.1 Pro Preview (API, high requested)	192/192	8/4	192/192	7/4	API

B.4 Additional conditional diagnostics

The following details complement the compact main-text comparisons. They use the same recorded responses and do not constitute additional experiments.

Recognition, task form, and appearance. Qwen3.8-27B identifies all 2,304 colors and 1,920 scored shapes, while 1,732/2,304 T1–T2 choices change (75.17%). Mistral changes 2,140/2,304 (92.88%). Gemini changes 130/2,304 (5.64%); GPT-6 Sol changes 4.69% with 99.90% joint accuracy versus Gemini's 100%. In the recognition-conditioned four-view analysis, Qwen3.8-27B disagrees in 259/576 scenes (44.97%), GPT-6 Sol in 194/574 (33.80%), and Gemini in 109/576 (18.92%). Qwen also disagrees across colors in 96/576 complete groups (16.67%); 77/480 fully correct, shape-eligible groups still disagree (16.04%). These units differ: a view group contains four images of one scene, while a color group contains four scenes sharing geometry.

Option order. InternVL3.5-38B retains all 192 colors correct across T2/D4 while 112 choices change; joint totals remain 158/160, which does not guarantee identical item-level attribute outcomes. Mistral changes 78/192 choices; Astra, GPT-5.6 Sol, GPT-6 Sol, and Gemini change 21, 12, 9, and 11. Both-valid denominators and configuration qualifications remain in the panel tables.

Tool and layout contrasts. The M03–M05 tool summaries each use 72 valid pairs per view, and the M04 comparisons use 24. For InternVL3.5-38B under T1, 549/1,152 layout pairs change (47.66%): 277 *that*-to-*this* and 272 reverse transitions yield a net $+0.43$ pp. Its addressee-eye and Exo contrasts are -77.78 and $+63.54$ pp. These T1 quantities are distinct from the main table's regional absolute T2 Exo magnitude. For Gemini under T2, bare-hand R2 layout contrasts are 0, -83.33, -4.17, and -29.17 pp in speaker eye, addressee eye, Exo, and Top, respectively, with 24 valid pairs per view. The external human R2 contrast is -4.69 pp with a 95% interval of $[-12.50, 3.13]$, which establishes neither a zero effect nor equivalence.

Conditional human-reference comparisons. The distance–layout reference declines from 65.10–72.40% *this* in R1 to 9.38–9.90% in R3 across layouts. Gemini selects *this* on every R1 Exo image and none in R3, with 0–4.17% in R2. Qwen3.8-27B's R2 rates are 37.50% opposite versus 4.17% side-by-side. Against the tool reference, Qwen3.8 Flash FP8 and GPT-6 Sol have pooled-view JSD 0.140463 and 0.140509; these are approximately equal at the four-decimal precision shown in Figure 4d. Against the distance–layout reference, Qwen3.8-27B's pooled JSD is 0.133, versus 0.400, 0.307, 0.231, and 0.347 in speaker eye, addressee eye, Exo, and Top. Its mean absolute error is 9.72 pp pooled versus 16.32–31.94 pp per view. These pooled-view scores are not the Exo-only main-table scores.

C Distance profiles for all deployments

All panels use common scales and twelve positions, with 192 expected images per position/task. Each curve pools tools, layouts, appearances, and views; coverage remains in the associated tables. Open weights are ordered by recorded total/component size, with API sizes undisclosed. D1 for the four early deployments crosses hardware/configuration boundaries. These curves are generated choice frequencies, not human-alignment scores.

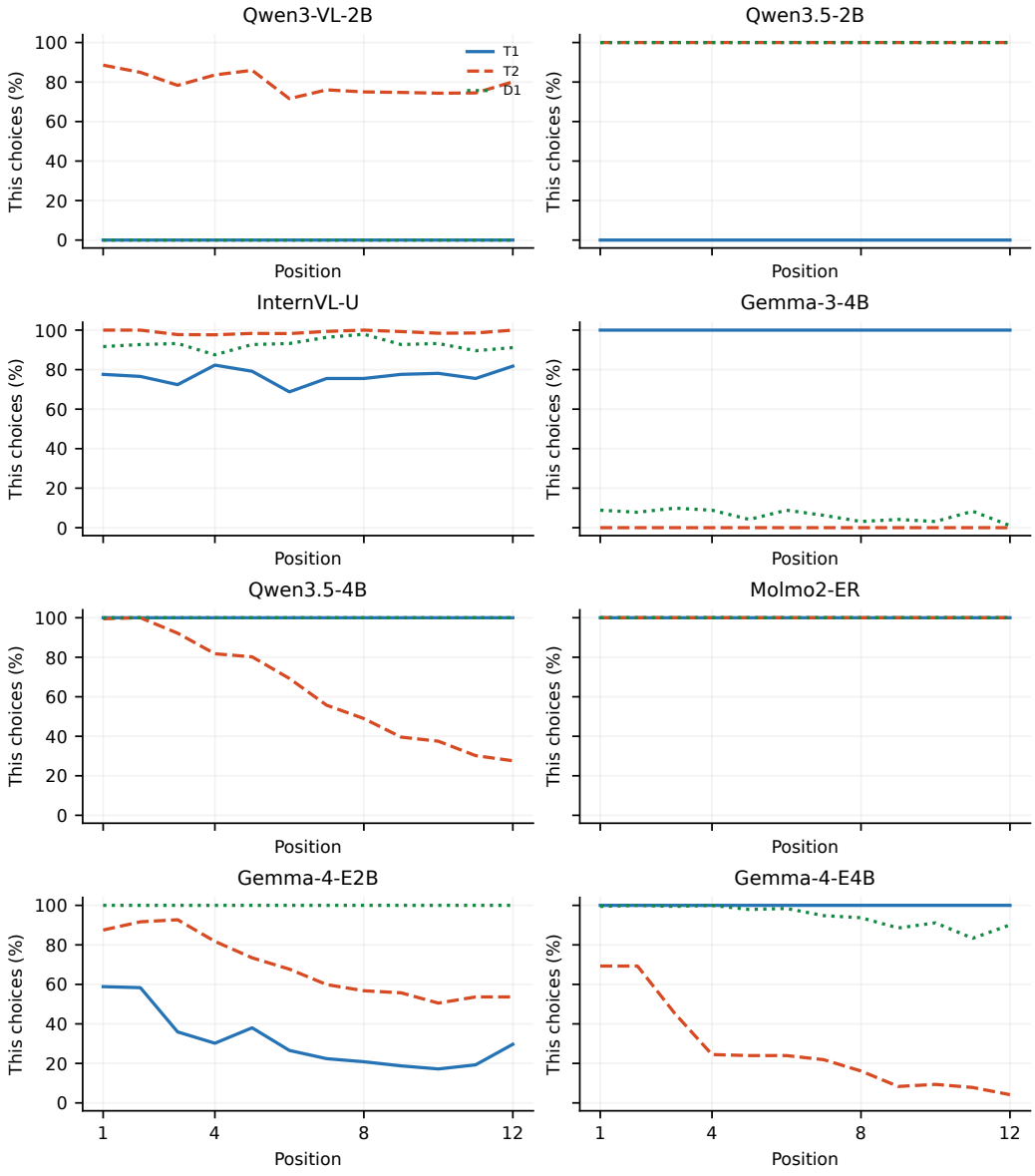


Figure 8: Complete distance profiles, census panels 1–8. T1 solid, T2 dashed, D1 dotted.

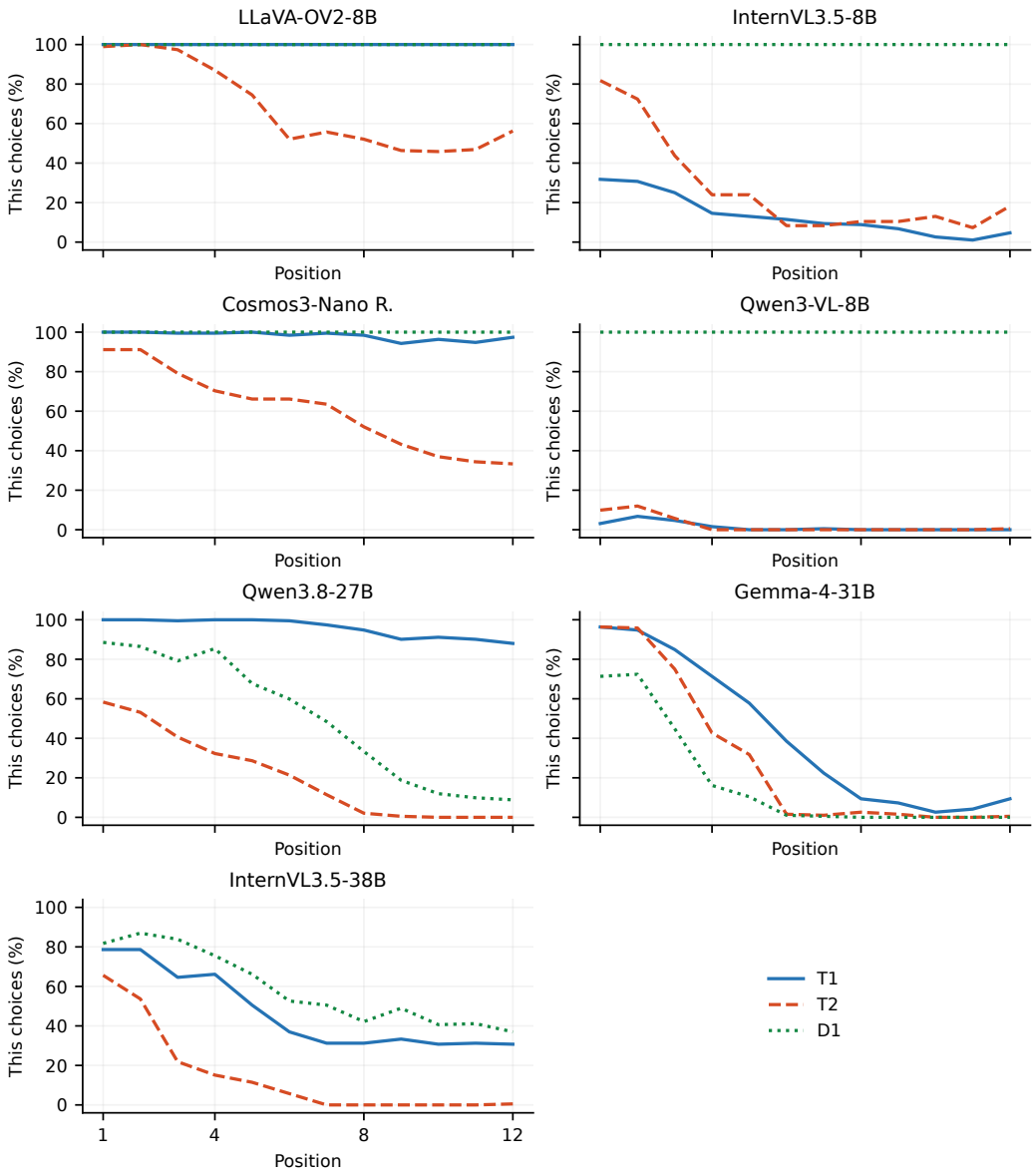


Figure 9: Complete distance profiles, census panels 9–15. T1 solid, T2 dashed, D1 dotted.

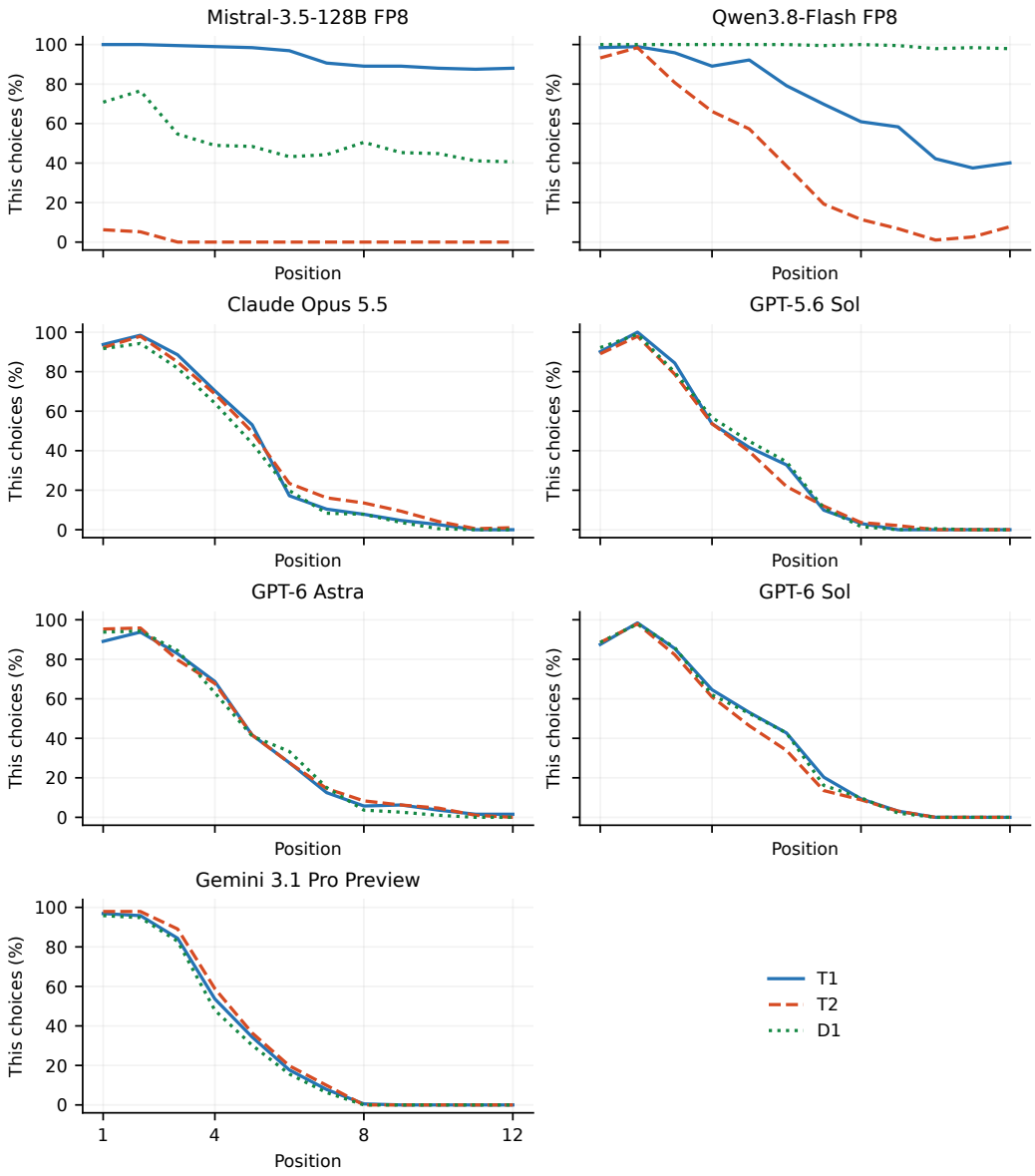


Figure 10: Complete distance profiles, census panels 16–22. T1 solid, T2 dashed, D1 dotted.

D EVALUATION PROMPTS

The following seven user prompts specify the response vocabulary, participant roles, and requested viewpoint. Their wording is reproduced exactly. Model-specific chat formatting is applied separately and is not part of these user messages.

D.1 T0: ATTRIBUTES ONLY

```
Identify the single three-dimensional object placed on the tabletop.
Ignore the small flat position markers and the humanoids.
Return exactly two lowercase words:  its color followed by its
three-dimensional shape.
Choose the color from:  blue, red, green, yellow.
Choose the shape from:  sphere, cube, cylinder.
Return only the two words, separated by one space.
```

D.2 T1: DEMONSTRATIVE ONLY

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
The speaker is referring to the colored three-dimensional object on the
tabletop.
Answer from the speaker's position, regardless of the camera viewpoint.
Complete the speaker's expression by choosing This or That.
Return the chosen word only.
```

D.3 T2: REFERRING PHRASE

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
The speaker is referring to the colored three-dimensional object on the
tabletop.
Answer from the speaker's position, regardless of the camera viewpoint.
Write the speaker's referring phrase as exactly three words in this
order:  demonstrative, color, three-dimensional shape.
Choose the demonstrative from:  this, that.
Choose the color from:  blue, red, green, yellow.
Choose the three-dimensional shape from:  sphere, cube, cylinder.
Return exactly the three chosen words, separated by single spaces, with
no punctuation or explanation.
```

D.4 D1: REVERSED DEMONSTRATIVE ORDER IN T1

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
The speaker is referring to the colored three-dimensional object on the
tabletop.
Answer from the speaker's position, regardless of the camera viewpoint.
Complete the speaker's expression by choosing That or This.
Return the chosen word only.
```

D.5 D2: ATTRIBUTE RECOGNITION WITH PARTICIPANT ROLES

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
Identify the single three-dimensional object placed on the tabletop.
Ignore the small flat position markers and the humanoids.
Return exactly two lowercase words:  its color followed by its
three-dimensional shape.
Choose the color from:  blue, red, green, yellow.
Choose the shape from:  sphere, cube, cylinder.
Return only the two words, separated by one space.
```

D.6 D3: ATTRIBUTES WITH FULL SCENE CONTEXT

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
The speaker is referring to the colored three-dimensional object on the
tabletop.
Answer from the speaker's position, regardless of the camera viewpoint.
Write the speaker's referring phrase as exactly two words in this order:
color, three-dimensional shape.
Choose the color from:  blue, red, green, yellow.
Choose the three-dimensional shape from:  sphere, cube, cylinder.
Return exactly the two chosen words, separated by single spaces, with no
punctuation or explanation.
```

D.7 D4: REVERSED DEMONSTRATIVE ORDER IN T2

```
The green-and-yellow humanoid (MIMo) is speaking to the black-and-white
humanoid (iCub).
The speaker is referring to the colored three-dimensional object on the
tabletop.
Answer from the speaker's position, regardless of the camera viewpoint.
Write the speaker's referring phrase as exactly three words in this
order:  demonstrative, color, three-dimensional shape.
Choose the demonstrative from:  that, this.
Choose the color from:  blue, red, green, yellow.
Choose the three-dimensional shape from:  sphere, cube, cylinder.
Return exactly the three chosen words, separated by single spaces, with
no punctuation or explanation.
```

E EXTERNAL HUMAN COMPARISON: DESIGN AND ESTIMANDS

These exploratory comparisons use external human response distributions, not human labels for the benchmark images. The correspondence uses a 0.6 distance scale, motivated by the smaller robot body: the compiled MIMo default-pose foot-to-head extent is 1.0408586 m. Dividing by 0.6 implies 1.734764 m, a design reference rather than a measured adult sample mean. A 700-mm human tool would scale to 420 mm, whereas the depicted robot rod is 520 mm; the construction does not assert uniform scaling or action equivalence. Final static contact bands likewise do not define the human regions.

E.1 COVENTRY 2008: PUBLISHED MEANS

The English Table 1 has three actor/placement-history means for each of two tools and four quadrants, giving 24 published rounded means. The reference used here averages the three actor means equally, producing eight cells. It does not recover exact twelve-position data, participant counts per cell, or uncertainty of the derived mean. Study-level English $N = 48$ is not substituted as each cell's sample size. Human Q1–Q4 map to M01–03, M04–06, M07–09, and M10–12, respectively. Models pool layouts and appearances, either within a view or across all views. This aggregation is a declared design correspondence, not a replication of placement history.

E.2 COVENTRY 2023: ENGLISH TRIAL DATA

The human trial data are restricted to English, yielding 32 participant identities and 1,152 rows. Each participant has 36 trials: two layouts, six distance codes, and three repetitions. Identical-looking rows are retained because they can represent legitimate repeated trials. The source has no tool manipulation; the primary model comparison therefore uses bare-hand images. Distance codes 1–6 map to M01, M02, M06, M07, M11, M12; region pairs define R1, R2, and R3. Intermediate positions have no direct reference. We compute $P(\text{This} \mid \text{region,layout})$ from counts, rather than transposing a plot of $P(\text{region} \mid \text{term})$.

Side-by-side This counts are 139, 33, and 18 out of 192 responses per region; opposite counts are 125, 42, and 19. Each cell includes all 32 participants. Pointwise 95% intervals use 10,000 bootstrap resamples of entire participant response vectors, using a fixed random seed, preserving within-person distance/layout dependence. The near–far contrasts are 63.02 pp side-by-side and 55.21 pp opposite. The R2 layout contrast is -4.69 pp with interval $[-12.50, 3.13]$; crossing zero is not evidence of equivalence.

E.3 Distribution distance and signed contrasts

For each mapped condition c, define $P_c = (p_c, 1-p_c)$ and $H_c = (h_c, 1-h_c)$ and $U_c = (P_c + H_c)/2$. We use

$$J_c = \sqrt{\tfrac{1}{2}\mathrm{KL}_2(P_c \| U_c) + \tfrac{1}{2}\mathrm{KL}_2(H_c \| U_c)}, \qquad \overline{J} = |C|^{-1} \sum_c J_c. \tag{2}$$

Thus J_c is JS distance, not divergence; the logarithm base is two. Absolute error is $100|p_c - h_c|$ in pp, averaged over the same conditions. The 2008 comparison has eight tool/quadrant cells; the 2023 primary comparison has six layout/region cells, with twelve layout/distance cells as a granularity sensitivity analysis. The model proportions use valid generated outputs and report expected counts, valid counts, and covered cells. Invalids are not reallocated as That. This estimand differs from candidate-sequence probabilities and from analyses conditioned on recognizing attributes. Human comparisons here use the unfiltered population of expected inputs as their coverage frame.

Signed held-minus-bare and side-by-side-minus-opposite contrasts are computed on jointly valid matched model pairs. Human contrasts retain their own design and uncertainty. Neither the distribution distance nor their contrast agreement is a current-image accuracy score. Full region/distance cells, per-view and pooled distances, near–far contrasts, and model-specific coverage are included with the analysis tables. No combined most-human ranking is defined.

F Summary metrics and complete decompositions

Distance trend. For layout ℓ, let $p_{m\ell}$ be T2 bare-hand Exo This frequency at each of the twelve distances $d_m = 0.15m$ metres, among valid outputs. We use an equal-position least-squares slope $\beta_\ell = \sum_m (d_m - \bar{d})(p_{m\ell} - \bar{p}_\ell) / \sum_m (d_m - \bar{d})^2$ and report $-100(\beta_{\text{opposite}} + \beta_{\text{side}})/2$ in pp/m. Positive values indicate a decreasing linear trend, not correctness. Every position contributes; nonlinear boundaries and reversals remain visible in the curves. Valid counts are retained for every position/layout cell. All registered items have answers, but unparseable outputs can change the contributing appearance mix. Their valid counts are reported separately; no choices are imputed.

Regional magnitudes. Q1–Q4 are consecutive groups of three markers. Within each quadrant and layout we form held-minus-bare matched T2 Exo differences; within each quadrant and tool state we form side-by-side-minus-opposite differences. Using the fractional contrast Δ from Equation 1, the main-table magnitude is $\frac{100}{8} \sum_{q=1}^{4} \sum_{s=1}^{2} |\Delta_{qs}|$ in pp, with 36 expected pairs per cell. This prevents between-cell cancellation but still permits opposite changes within a cell to cancel. A large magnitude does not establish a human-like sign or location. The six-anchor R1–R3 mapping is reserved for the 2023 human comparison and does not restrict these full-pool context summaries.

View, task, and order. View is $100N_{\text{mixed}}/N_{\text{complete}}$ for four-view T2 scene groups; all four strict-valid answers must agree to count as stable. The six pairwise view rates diagnose which observation changes matter and are distinct from the four-way criterion. Task form compares same-image T1/T2; option order compares T2/D4 on the fixed panel. API single-pair differences can include stochastic or deployment effects; constant outputs yield zero changes without demonstrating contextual use.

Human distance and overlap. The primary human score is the arithmetic mean of six condition-wise, square-root base-2 JS distances for T2 bare-hand Exo across three regions and two layouts. This adopts the distribution-distance form used in Watanabe et al. (2026); generated response frequencies and our all-strict-valid selection differ from a logit-based, attribute-filtered protocol. The absolute-error sensitivity check is $100[1 - \frac{1}{6}\sum_c |p_c - h_c|]$, equal to mean distribution overlap for

binary This/That. Neither is current-image accuracy. Human self-comparison is mathematically zero distance, not an independently measured human reliability baseline.

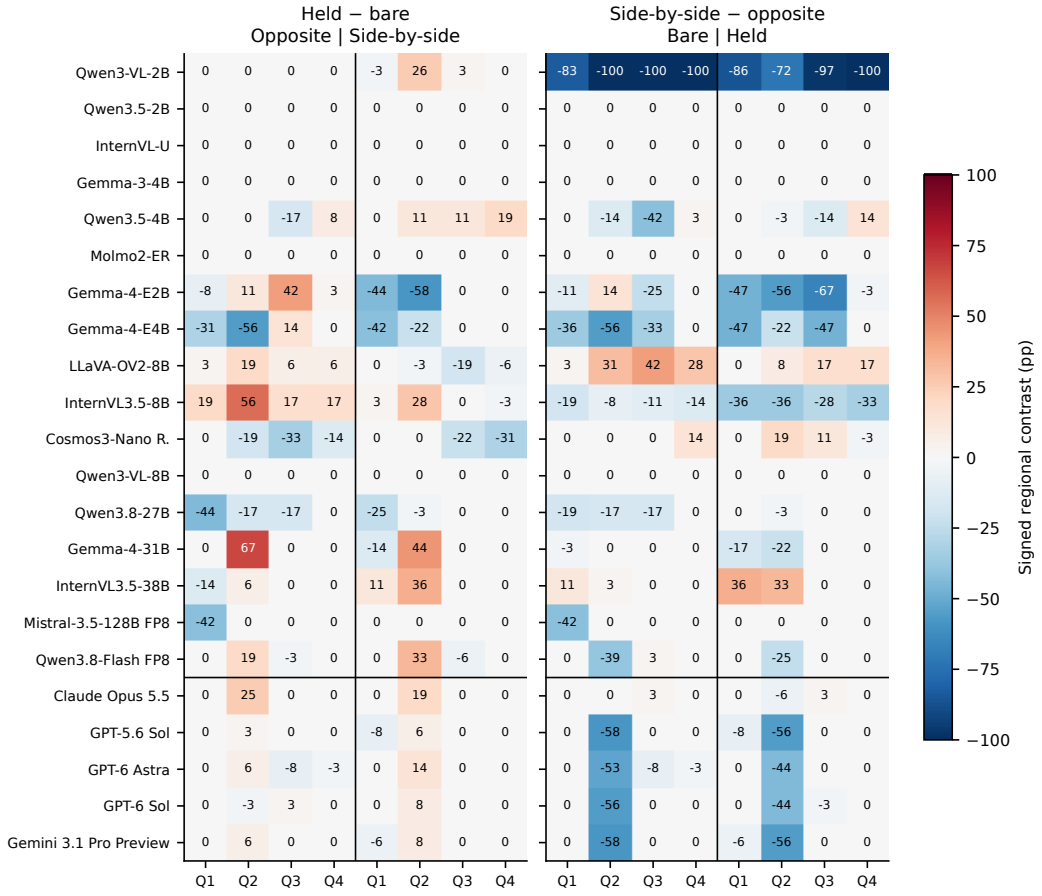


Figure 11: All 22 deployments: signed T2 Exo contrasts in four quadrants, separated by the other scene factor. The main-table Tool/Layout values average the absolute values of their respective eight cells. Color shows signed pp; exact valid denominators accompany the data tables.

Reading the main table. All registered runs are complete; the reported valid counts describe whether outputs meet the strict response contract. Boldface compares models within the API and open-weight groups using unrounded values. Attribute maxima use the expected applicable denominator. JSD and sensitivity minima require the full expected valid comparison set; Order also requires matched deployment configurations. Lower sensitivity means fewer observed changes, and context magnitudes have no universal better direction. Human and analytic-reference rows are excluded from model-group extrema.

Matched human-reference summaries. The external human row of Table 1 reports study-specific summaries. Matched model comparisons in this appendix recompute responses on the corresponding mapped support rather than reusing the full-benchmark contextual columns. For the 2023 distance slope, both human and model proportions use M01, M02, M06, M07, M11 and M12, expressed in robot-coordinate metres under the stated scale mapping. Human means give 39.24 pp/m. The 2008 mean absolute four-quadrant tool contrast is 6.83 pp, using the published rounded means; model layouts are pooled within each quadrant before taking absolute values. The 2023 bare-hand mean absolute three-region layout contrast is 4.17 pp. These are descriptive point estimates from distinct studies, not a combined human benchmark score.

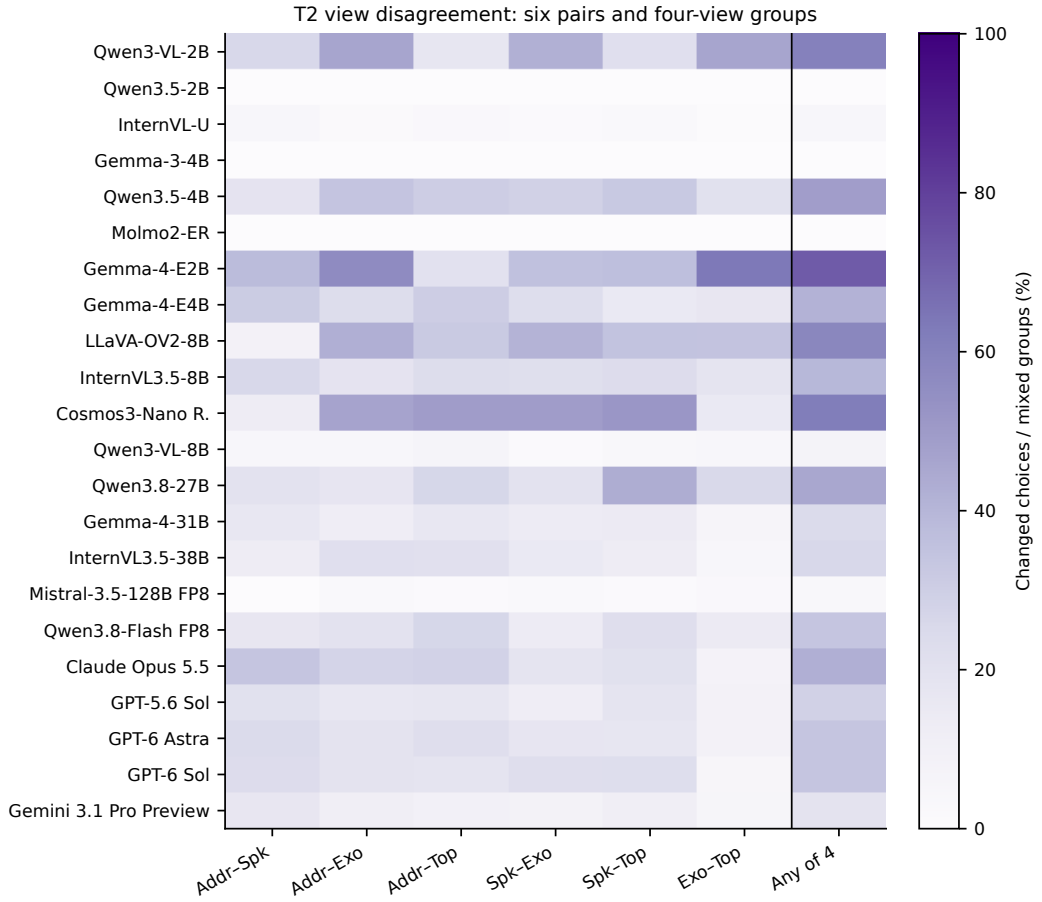


Figure 12: All six pairwise view comparisons and the four-view mixed-choice criterion. Each column uses its own valid pair/group denominator; four-view disagreement is not the mean of the six pair rates.

Table 11: Exact coverage of revised main-table summaries. Distance/Human count responses; Tool/Layout count paired comparisons; View counts complete four-view scene groups; Task/Order count jointly valid pairs. Different columns therefore have different units of coverage.

Model	Distance	Tool	Layout	View	Task	Order	Human
Claude Opus 5.5	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
GPT-5.6 Sol	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
GPT-6 Astra	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
GPT-6 Sol	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Gemini 3.1 Pro Preview	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Qwen3-VL-2B	287/288	287/288	287/288	564/576	2292/2304	179/192	143/144
Qwen3.5-2B	209/288	200/288	201/288	288/576	1595/2304	127/192	106/144
InternVL-U	205/288	173/288	131/288	263/576	1863/2304	130/192	101/144
Gemma-3-4B	277/288	271/288	270/288	483/576	2208/2304	179/192	136/144
Qwen3.5-4B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Molmo2-ER	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Gemma-4-E2B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Gemma-4-E4B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
LLaVA-OV2-8B	288/288	288/288	288/288	576/576	2290/2304	192/192	144/144
InternVL3.5-8B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Cosmos3-Nano R.	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Qwen3-VL-8B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Qwen3.8-27B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Gemma-4-31B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
InternVL3.5-38B	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Mistral-3.5-128B FP8	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144
Qwen3.8-Flash FP8	288/288	288/288	288/288	576/576	2304/2304	192/192	144/144

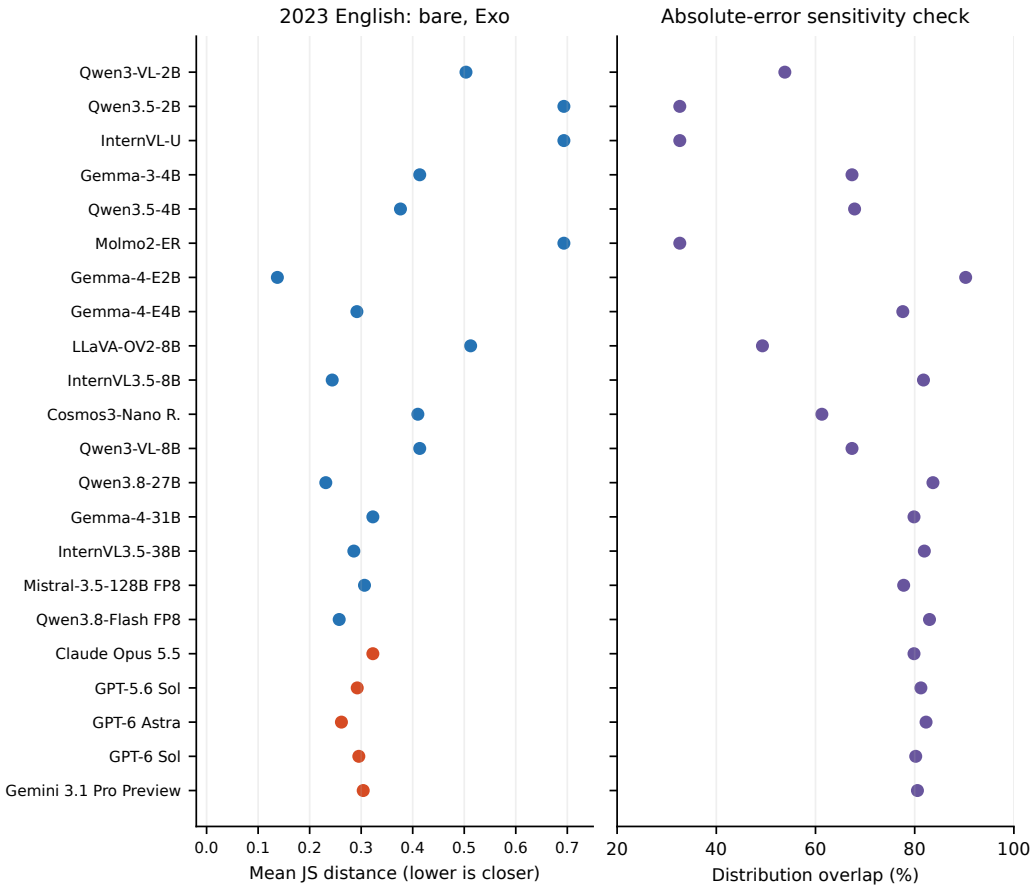


Figure 13: 2023 English reference for all 22 deployments, fixed to T2 bare-hand Exo: JS distance and the corresponding absolute-error-based distribution overlap. Both retain six region-by-layout cells. Overlap is a supplementary diagnostic; Table 1 uses JSD. Strict-valid counts are provided in the accompanying table.

The accompanying machine-readable group summaries use deployment medians over one fixed complete-coverage subset for all three matched metrics: all five APIs and 13 of 17 open-weight deployments. Every individual deployment, including incomplete comparisons, is retained in Table 12. This avoids comparing a full-twelve-position or two-tool model response with a narrower human design. The human studies provide no corresponding four-camera, T1/T2, option-order or current-image attribute task. Human-reference self-comparison would be identically zero JSD and is not treated as an independently measured reliability baseline.

Table 12: Context responses on matched external-reference support. All model summaries use T2 Exo. Distance uses six 2023 anchor positions in mapped robot-coordinate metres and averages layout-specific negative slopes; Tool uses four 2008 quadrants with model layouts pooled before taking absolute differences; Layout uses three 2023 regions in the bare condition. Tool/Layout average absolute region differences. Coverage lists Distance responses / Tool pairs / Layout pairs. These supports differ from the full-benchmark Context response columns in Table 1. *Incomplete coverage. External studies remain distinct.

Model/reference	Distance	Tool	Layout	Coverage
Coventry 2008	–	6.8	–	–
Coventry 2023 (English)	39.2	–	4.2	–
Proprietary (API)				
Claude Opus 5.5	65.7	5.6	0.0	144/144 / 288/288 / 72/72
GPT-5.6 Sol	66.0	2.1	2.8	144/144 / 288/288 / 72/72
GPT-6 Astra	65.3	3.8	12.5	144/144 / 288/288 / 72/72
GPT-6 Sol	67.0	1.0	13.9	144/144 / 288/288 / 72/72
Gemini 3.1 Pro Preview	65.8	2.4	1.4	144/144 / 288/288 / 72/72
Open-weight				
Qwen3-VL-2B	8.8*	3.9*	91.7*	143/144 / 287/288 / 71/72
Qwen3.5-2B	0.0*	0.0*	0.0*	106/144 / 200/288 / 52/72
InternVL-U	0.0*	0.0*	0.0*	101/144 / 173/288 / 34/72
Gemma-3-4B	0.0*	0.0*	0.0*	136/144 / 271/288 / 64/72
Qwen3.5-4B	64.6	5.6	16.7	144/144 / 288/288 / 72/72
Molmo2-ER	0.0	0.0	0.0	144/144 / 288/288 / 72/72
Gemma-4-E2B	48.6	18.1	6.9	144/144 / 288/288 / 72/72
Gemma-4-E4B	50.2	20.5	37.5	144/144 / 288/288 / 72/72
LLaVA-OV2-8B	6.0	4.2	22.2	144/144 / 288/288 / 72/72
InternVL3.5-8B	52.8	17.0	11.1	144/144 / 288/288 / 72/72
Cosmos3-Nano R.	55.6	14.9	5.6	144/144 / 288/288 / 72/72
Qwen3-VL-8B	0.0	0.0	0.0	144/144 / 288/288 / 72/72
Qwen3.8-27B	62.0	13.2	15.3	144/144 / 288/288 / 72/72
Gemma-4-31B	65.7	15.6	0.0	144/144 / 288/288 / 72/72
InternVL3.5-38B	61.2	5.6	4.2	144/144 / 288/288 / 72/72
Mistral-3.5-128B FP8	20.4	5.2	20.8	144/144 / 288/288 / 72/72
Qwen3.8-Flash FP8	66.5	7.6	8.3	144/144 / 288/288 / 72/72

Analytic random baseline. Each image–prompt item independently samples a demonstrative from two options, a color from four, and a shape from three, and emits a strict-valid answer. Expected color and joint-attribute accuracy are 25% and $100/12\%$. The signed distance slope has expectation zero. Four independent views disagree with probability $1 - 2(1/2)^4 = 0.875$; Task and Order each change with probability $1/2$. For Tool/Layout, each cell compares two independent sample proportions with $n = 36$. Their expected absolute difference in pp is $100\binom{2n}{n}/4^n$, not zero. Human JSD is the mean of six expected sample distances: in each cell, average the base-2 JS distance between $[K/24, 1 - K/24]$ and the human distribution over $K \sim \mathrm{Binomial}(24, 1/2)$. Thus the baseline accounts for the finite sample sizes of the reported statistics. It involves no new model calls and is excluded from the 22-deployment count and model-group bolding. The compact Human reference row combines separately sourced summaries: 2023 distance/layout and 2008 tool contrasts; it does not combine participants or imply a shared experimental protocol.

G Appearance specifications and reference renders

G.1 Appearance family and reference renders

The formal collection fully crosses three shapes and four colors at every position and layout (Section 3). Figure 14 retains the twelve historical standalone appearance references. These are separate static object renders, not formal robot-eye images, a multi-target scene or additional physical trials. Their lighting and material settings describe this reference figure only; formal scene materials and cameras are independently registered.

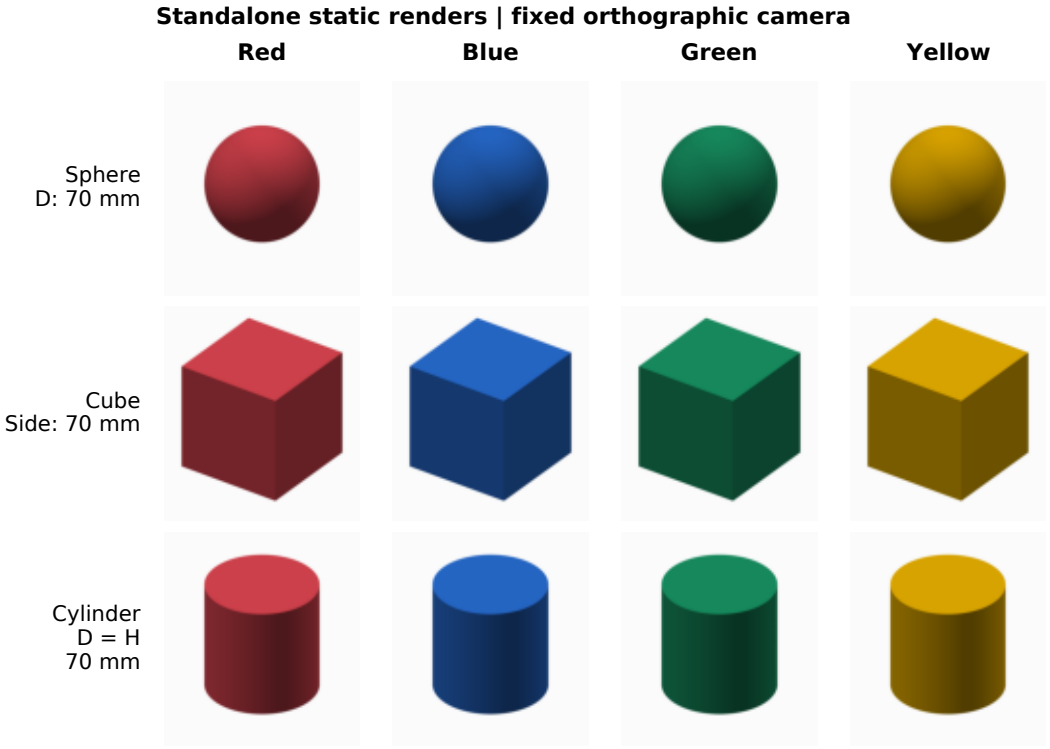


Figure 14: Complete twelve-appearance reference: three shapes crossed with four colors (D: diameter; H: height). Each panel reuses an existing standalone MuJoCo 3.3.0 static render, with common orthographic camera, scale, matte material settings and lighting. All panels receive the same blank-background crop and display scale; object colors and silhouettes are unchanged. This comparison is not a simultaneous twelve-object scene, a robot-eye view, a reachability witness, or a recognition test. It is distinct from the frozen physical-evidence renders in Figures 17 and 20.

Table 13: Primitive dimensions and orientation (shared with formal construction). The common axis-aligned bounding box is $70 \times 70 \times 70$ mm; volume and projected area are not matched.

Shape	Dimensions (mm)	Orientation	Volume (cm^3)
Sphere	Diameter 70	Rotation invariant	179.594
Cube	Side 70	Table-axis aligned; yaw 0°	343.000
Cylinder	Diameter 70; height 70	Axis parallel to world $+Z$	269.392

G.2 Scale, color and rendering controls

The sphere diameter, cube side, and cylinder diameter and height are all 70 mm. The cube follows the table axes; the cylinder axis is vertical. Their common bounding envelope does not equalize

volume (Table 13) or projected area. Shape-dependent image extents are retained. Objects are not enlarged with distance. These are simulation implementation choices, not a reconstruction of the shapes, colors or scale in the human experiment (Coventry et al., 2008). The pilot's historical 56-mm sphere is also a separate geometry condition and supplies no validation of the 70-mm family.

Table 14: Manifest-defined material colors, used for every shape. RGB values are normalized channel codes before float32 storage, with alpha 1; they are not measured pixel colors.

Color	Requested hex	Material RGB input
Red	`#C83F49`	(0.784314, 0.247059, 0.286275)
Blue	`#2463BF`	(0.141176, 0.388235, 0.749020)
Green	`#16865B`	(0.086275, 0.525490, 0.356863)
Yellow	`#D4A000`	(0.831373, 0.627451, 0.000000)

Hex channel codes in Table 14 are divided by 255 and passed directly to MuJoCo material RGBA; no sRGB-to-linear transfer is applied to those inputs. The material metadata additionally record the compiled float32 values and separate linear-RGB reference values. Lighting changes the rendered pixels, so these settings are not a colorimetric calibration or a measured discriminability guarantee. Across all twelve references, alpha is 1, specular, shininess and reflectance are zero, and textures, headlight and cast shadows are disabled.

The fixed overview camera is orthographic, at $(0.1875, -0.2625, 0.225)$ m, aimed at $(0, 0, 0.0325)$ m, with a 0.130-m vertical view extent. Each source tile is 1008×620 pixels. The key light has position $(-0.18, -0.25, 0.45)$ m, direction $(0.3, 0.4, -1)$, diffuse intensity $(0.55, 0.55, 0.55)$ and ambient intensity $(0.30, 0.30, 0.30)$. The fill uses $(0.25, 0.10, 0.28)$ m, direction $(-0.6, -1, -0.7)$, diffuse $(0.28, 0.28, 0.28)$ and ambient $(0.08, 0.08, 0.08)$. These common reference settings document the appearance overview; they do not establish a final robot-eye input pipeline.

G.3 Placement and historical projection estimates

Placement checks use the compiled table geometry and saved poses. The table has center $(1.075, 0, 0.400)$ m and half-extents $(0.975, 0.4875, 0.025)$ m. Marker XY centers lie on $y = 0$, from $x = 0.250$ to 1.900 m. Table 15 distinguishes the standalone overview origin from current tabletop placement. The physical top, not the raised visual marker, defines the new object's bottom height.

Table 15: Position anchors and recorded world-coordinate placement checks for the 70-mm family. Placement arithmetic does not establish dynamic feasibility.

Item	Specification or audit result
Standalone reference base / center	$(0, 0, 0)$ / $(0, 0, 0.035)$ m
Formal target base center	$(x_i, 0, 0.425)$ m; $x_i = 0.100 + 0.150i$ m, $i = 1, \ldots, 12$
Formal object center	$(x_i, 0, 0.460)$ m for all three shapes
Visual marker center height	$z = 0.42515$ m; not the physical support height
Adjacent marker spacing	150 mm
Adjacent aligned footprint gap	$150 - 70 = 80$ mm
Minimum footprint–table boundary gap	115 mm, over the twelve placements

The 80-mm interval clearance and in-table footprints are world-coordinate placement results. They do not certify separation from cloth, robot bodies, or other scene geometry. Nor do they prove separation in an eye image: perspective can make front objects occlude rear ones. The formal collection uses one target per image. The standalone overview is not a multiple-target input; any future simultaneous-object condition would require separate visibility and target-selection checks.

Table 16 is a historical estimate, not an extent measurement of the formal images. It uses archived gaze frames with a hypothetical 1600×1100 raster and vertical fields of view of 60° for MIMo's

Table 16: Historical pilot only; not measured extents of benchmark images. M12 geometric image-extents for 70-mm objects at 1600×1100 pixels. These are unoccluded projections, not rendered segmentation masks or recognition measurements.

Shape	MIMo left eye, 60° FOV Width $\times$ height (pixels)	iCub midpoint, 80° FOV Width $\times$ height (pixels)
Sphere	37.1×37.1	23.4×23.4
Cube	40.2×41.3	26.7×26.7
Cylinder	37.2×41.0	23.5×26.3

left eye and 80° for the iCub midpoint proxy. The latter is not either individual eye. Sphere extents are analytic, cube extents use projected vertices, and cylinder rims are sampled every 0.5°. These geometric estimates ignore occluders; the cylinder sampling is not a certified continuous-silhouette bound. At M12 the iCub estimates are only about 23–27 pixels per dimension. This records scale, not a passed vision-language-model (VLM) recognition test. The current native formal raster is 1920×1080 RGB with native left eyes and 60° Ego display fields of view. Those historical projection estimates do not test model-specific preprocessing or recognition; the later formal-pool VLM evaluation is reported separately. None of the historical estimates supplies an exact current-image extent.

G.4 Evidence scope after formal construction

The standalone reference set contains twelve appearance images with specified geometry, materials, and placement. The benchmark collection separately contains 576 scenes with static geometry checks across its shape and color conditions (Appendix H). Color variants reuse physical states. Successful geometry checks do not remove the documented image-extent discrepancies or establish human or model recognition.

The earlier radius-7.5-mm disk studies, dynamic trace, and conditional certificate retain their original geometry and configuration domains. The bare-hand M03 and rod-assisted M06 blue-sphere contact examples are separate from the benchmark's non-contact pointing states. The benchmark scenes have not undergone dynamic validation, and image-level human annotations are unavailable. Model results concern the benchmark image collection, not the historical reference renders or hypothetical projection estimates.

H FORMAL CONSTRUCTION: GEOMETRY, ADAPTATION AND LAYOUT

H.1 SCENE STRUCTURE AND GEOMETRIC REPRESENTATION

The benchmark uses a fixed collection of simulated scenes. Each tool condition comprises 72 geometry templates and 288 color variants. A template combines tool condition, layout, position, and shape; a scene adds color, and an image adds viewpoint. The collection retains 192 previously rendered blue-sphere images. Model evaluation and subsequent analysis use these saved stimuli without altering their geometry or rendering new inputs.

Sphere, box and upright cylinder dimensions are assigned to the corresponding geometric primitives. Visual and physical shape are determined from these primitives rather than internal object names. Both visual and collision queries use the registered primitive. The bare-hand pipeline synchronizes type, size, radius bound, axis-aligned bound and the target's bounding-volume-hierarchy leaf where present. The tool model keeps the target as its existing single geometry without a bounding-volume hierarchy and updates its primitive bounds. Source validators independently reconstruct the allowed model changes and compare compiled arrays before checking the saved pose. These static checks do not establish any time integration, stable force-bearing contact, balance or human assessment.

H.2 SEPARATE ADAPTATION AND ACCEPTANCE RULES

For bare hand, the legal contacting feature is the forward distal index endcap, not any nearby hand surface. The recorded closest point must lie on that endcap, with nonnegative surface gap at most 0.05 mm and whole-capsule penetration no greater than 10^{-8} m. Whole-capsule and endcap distances must agree within 10^{-7} m. All 21 right-hand geometries are included in the hand checks; non-tip hand–target and mask-independent hand–solid-table clearances must be at least 3 mm (with the source's 10^{-9}-m numerical allowance). Native joint/tendon acceptance retains 10^{-5}-rad tolerances. Registered support and abnormal-penetration checks remain distinct from target contact.

Only a failed contact template may receive local shoulder, elbow or wrist adaptation, bounded by 12° per registered coordinate relative to its baseline. All 22 right-finger coordinates and nonregistered state coordinates are preserved. If a new silhouette violates the inherited 3% image margin, necessary head/eye adjustment is limited to 5° and logged separately. Eight bare-hand templates change pose; two of these, M01 cubes in the two layouts, also change head/eye coordinates. No torso changes are recorded. For a ray state, the whole-capsule gap must be positive, the center projection along the actual index axis must be forward, and the perpendicular center error obeys $e_\perp \leq 0.001\ell + 10^{-6}$ m for forward distance ℓ in metres. Actual primitive intersection and first-occluder checks remain separate.

For held tool, the designated effector is the complete 5-mm-radius physical tip sphere. The verifier queries the target's true signed surface distance from the tip center and subtracts tip radius; it retains nearest surface points and face, edge or rim features and cross-checks native geometry distance. Contact accepts a gap of at most 0.10 mm with at most 1 μm numerical penetration. Forward-axis center error is limited to 1 mm, with forward first intersection and occlusion separately recorded. This fixed distance criterion differs from the bare-hand relative ray criterion. The whole shaft, environment, geometric grip, left-hand support, both robots and native constraints retain their registered checks.

The tool pipeline substitutes shape first. A failed contact template may slide the rod axially while preserving the 22 grasping-finger coordinates and palm-relative rod rotation and transverse placement. Fifteen tool templates record changed pose coordinates; all retain head/eye and torso coordinates. These counts describe deterministic adaptations to reused baselines, not independent trials. The fingertips' contact locations, axial forward point and pointing-ray intersection are not forced to be the same point. All acceptance tolerances are engineering conventions. Neither the selected contact range nor a saved positive gap certifies unreachability of an alternative configuration.

H.3 COORDINATES, SEATING AND RETAINED CLOTH

The solid table spans $x = 0.100$ to 2.050 m, with top at $z = 0.425$ m. In side-by-side layouts MIMo's hip-anchor midpoint is near $y = 0.130$ m and iCub's near $y = -0.300$ m; both sit at the

near short end. Their recorded baseline seat distances are approximately 150.0017/150.1798 mm for bare hand and 150.0000/150.0661 mm for held tool (MIMo/iCub). In opposite layouts the registered anchor midpoints are centered at $y = 0$, with $x = -0.050$ m for MIMo and $x = 2.200$ m for iCub, giving 150 mm from each robot's own edge. Seating uses actual bilateral hip-flexion anchors, not pelvis body origins. The bare-hand opposite validator checks this layout independently within 10^{-6} m.

Seat distance is measured from each robot's own table edge. A signed coordinate difference from the speaker's near edge is not an addressee seat distance; in the opposite layout, this distinction prevents confusing a world-coordinate offset with the prescribed 150-mm seating separation.

The cloth treatment is layout-specific. Side-by-side scenes retain their registered baseline drape. Opposite-layout registrations keep the original top faces and both long-side folds while opening the entire short ends, including the corresponding corner halves. They do not create a local robot-shaped window. The two pipelines preserve the retained surface but differ in their stored acceleration structures; these implementation details are not visual factors. The top cloth lies about 25 μm above the physical table, so the reported target-bottom/visual-cloth overlay is retained. Tangency to the physical table and cloth checks are different claims. No cloth, rod, hand or occluder was removed during manuscript figure assembly.

I Formal views and annotation limits

Native left eyes supply the formal Ego inputs and native right eyes supply audit images. Origins, forward/up axes, projection type, field of view, and orthographic scale are stored per image. The 60° Ego value is a display setting, not a calibrated biological optical model. Exo uses the inherited fixed camera per layout; Top preserves its orthographic camera. Figures 15 and 16 show all formal views of the same M04 blue-sphere scenes. Columns separate layouts and rows separate views to retain readable panels. Each image is shown in full. They are four alternative single-image inputs per scene, not a multi-image prompt.

Pixel diagnostics include visible target segmentation, shape-aware isolated-ray visibility, center visibility, projected bbox/margins, foreground versus surrounding-ring RGB difference, and a proportional 768-pixel-width diagnostic. They do not test recognition after a model's actual resize/crop pipeline. The Top sphere/cylinder rule is an a priori applicability restriction on exact three-dimensional shape scoring, not a change in object identity or a claim that every individual image is indistinguishable. All-view diagnostics preserve those 384 images and their factual labels.

Table 17: Recorded visible-versus-projected bbox discrepancies greater than 2 pixels. These are metadata findings, not a count of damaged RGB files.

Tool condition	View / role	Flagged records
held tool	Exo	144
no tool	Speaker right eye (audit)	144
no tool	Speaker left eye	144
no tool	Exo	192

The independent packaging audit compares the segmentation-derived visible bbox against the stored projected bbox and flags an outward discrepancy greater than 2 pixels on any side. Table 17 accounts for all 624 pending records. A source QA pass and a pending extent finding can coexist. The stored bbox is not a certified exact continuous silhouette; neither these records nor source pixel checks establish that every target is fully in frame or unobscured. No mask re-render was performed to resolve the findings.

The bare-hand perspective path delegates bounds to its inherited visibility routine, whereas the tool path records shape-specific primitive projections. The pair table records algorithm mismatches for 96 cylinder pairs in each perspective view, including Exo. That difference concerns metadata computation; it cannot establish that equal-size rendered cylinders differ in visual size. The complete native-image bbox fields, methods and flags remain in the frozen image tables.

J DATA ORGANIZATION AND COMPARISON UNITS

The data are organized into linked tables for geometry templates, scenes, evaluation images, supplementary right-eye images, quality findings, and matched comparisons. External human summaries and their mappings are stored separately. Scene and image identifiers link each observation to its geometry, camera, and experimental conditions. The package contains 15 tables with typed columnar representations; consistency checks compare table contents and preserve missing values across formats.

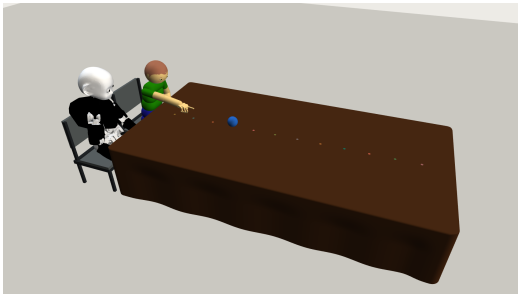
Lengths use metres, camera angles degrees and native hinge coordinates radians. Pose arrays require their original named-joint model and cannot be compared by index across models as if joint semantics were automatically identical. The three levels of annotation are factual state (shape, color, dimensions, coordinates, contact/ray and camera), protocol metadata (speaker, groups, applicability and prompt specification), and behavioral information (current-image responses or external summaries). The first two do not determine the third.

Table 18: Condition groups preserve their declared unit and dependencies; they do not create additional observations.

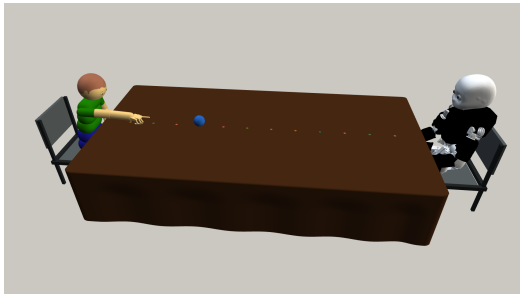
Comparison	Groups	Grouping unit
Tool	288	Two scenes; same layout, position, shape, color
Color	144	Four scenes sharing one geometry template
Shape	192	Three scenes at fixed tool, layout, position, color
Layout	288	Two scenes at fixed tool, position, shape, color
Viewpoint	576	Four images from one scene

Colors share one geometry template and physical camera parameters; views share one scene. The 24 dependency families group by layout and target position across tool and appearance variants. Both levels and the listed pairings matter for any later resampling or held-out analysis. No random split, bootstrap, significance test or new holdout was introduced. Complete-group coverage accompanies viewpoint agreement and color/shape sensitivity. Shape adaptation and different gaze policies are recorded residual changes, rather than silently assumed controlled factors.

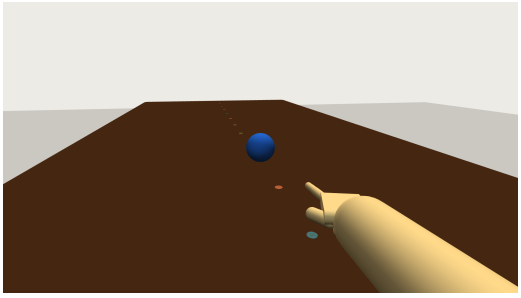
a Exo | Side by side



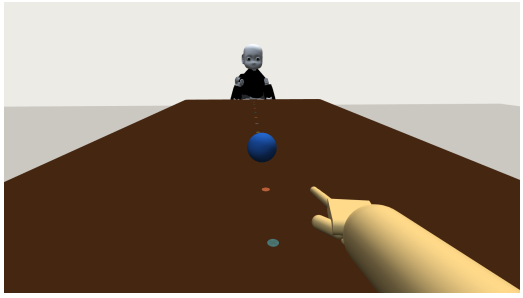
b Exo | Opposite



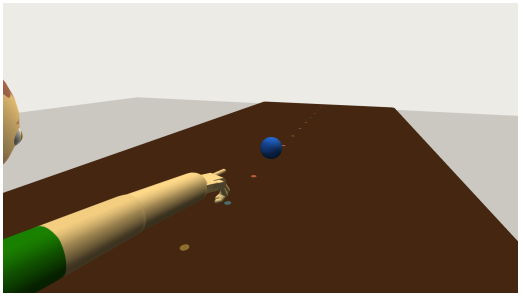
c Ego-MIMo (left) | Side by side



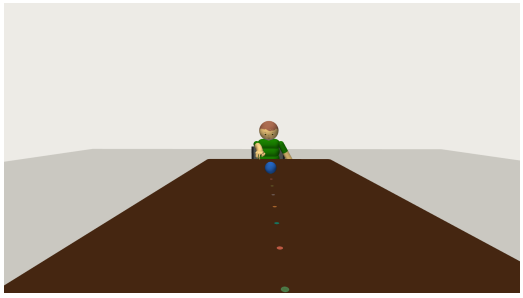
d Ego-MIMo (left) | Opposite



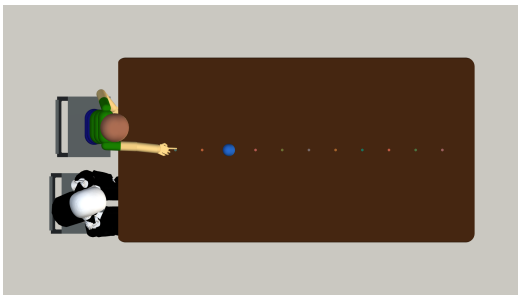
e Ego-iCub (left) | Side by side



f Ego-iCub (left) | Opposite



g Top | Side by side



h Top | Opposite

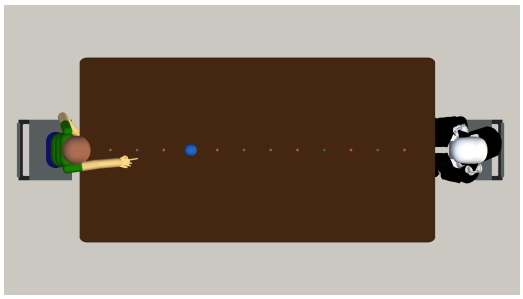
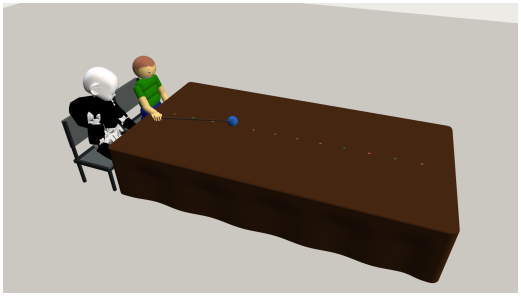
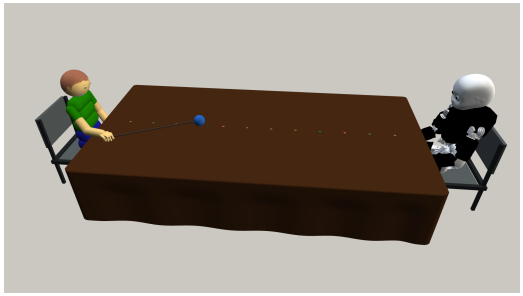


Figure 15: Bare-hand M04 blue sphere, four formal views in two layouts. All panels share their layout-specific saved scene. Original frame boundaries and occlusion remain visible. Changing viewpoint does not change the designated MIMo speaker. The pointing state is not an unreachability label.

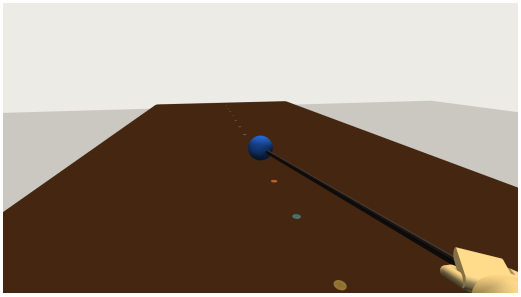
a Exo | Side by side



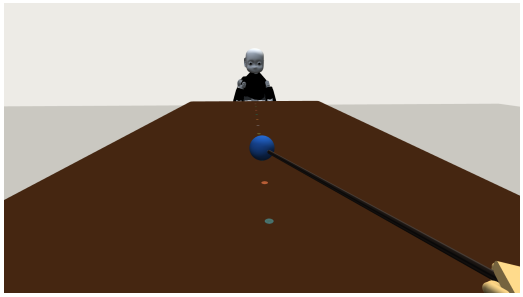
b Exo | Opposite



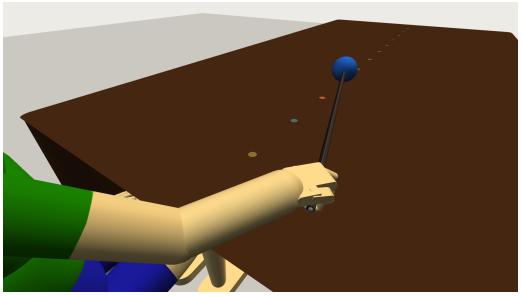
c Ego-MIMo (left) | Side by side



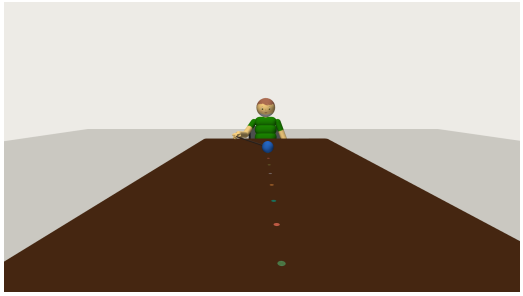
d Ego-MIMo (left) | Opposite



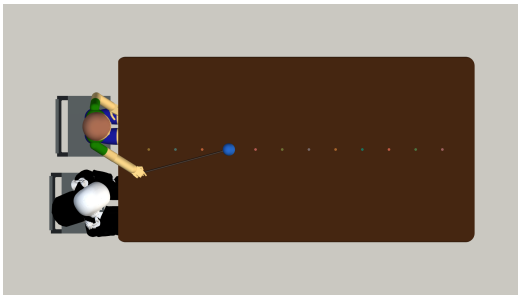
e Ego-iCub (left) | Side by side



f Ego-iCub (left) | Opposite



g Top | Side by side



h Top | Opposite

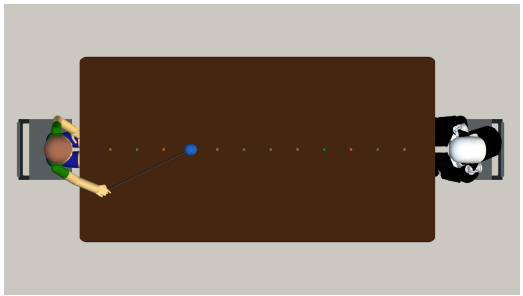


Figure 16: Held-tool M04 blue sphere, corresponding four formal views. Exo/Top cameras match the bare-hand condition; Ego cameras do not. The stored contact classification comes from geometric records, not visual measurement of the tip gap. All saved pixels are retained. No diagnostic close-up has been fabricated.

K Evaluation implementation

K.1 Exact prompts and parsing

Evaluation uses English prompts and one image per call. Model inputs contain only the image and prompt text; participant roles, target identity, and requested viewpoint are specified by the task. The T1 and T2 prompts are fixed; diagnostic variants introduce their specified wording changes.

The complete current task texts are provided in Appendix D.

The scorer strips outer whitespace, lowercases and splits on whitespace. T1 accepts one token from {this, that}; T2 accepts three tokens in order from {this, that}, {blue, red, green, yellow}, and {sphere, cube, cylinder}. It rejects punctuation, explanations and out-of-vocabulary tokens. The prompt requests single spaces; the implemented parser accepts whitespace-separated tokens, including repeated whitespace. A null or absent response is missing; a present unparsable response is invalid. Neither is converted into *that*. Duplicate, unknown and audit image IDs, mixed tasks and unsupported confidence fields are rejected by the response contract.

K.2 Metrics and denominators

Let N be the expected applicable image count, N_p the present non-null response count and N_v the valid count. The scorer reports expected, present, valid, missing and invalid counts, format validity N_v/N and N_v/N_p, and coverage. Demonstrative choice is the fraction of valid responses selecting *this*. It is not accuracy because no demonstrative gold labels exist.

For T2, color accuracy is reported both among valid responses and against all expected applicable images. Primary shape accuracy is reported among valid shape-eligible responses and against all 1,920 shape-eligible images. Missing and invalid responses do not enter the correct numerator, but remain in the all-eligible denominator. Empty denominators yield unavailable values, never an imputed zero. Full-view shape diagnostic accuracy and explicit shape coverage preserve the Top ambiguity restriction. Table 1 uses expected-denominator Color and Joint accuracy; Joint requires both color and the eligible shape to be correct. T1 returns no factual color/shape score. Deployment documentation specifies model configuration, chat formatting, tokenization, image resizing/cropping, candidate extraction, and generation settings (Appendix A).

For finite candidate-sequence log likelihoods $\ell_{\text{this}}, \ell_{\text{that}}$, the helper normalizes as $p_k = \exp(\ell_k - m)/\sum_j \exp(\ell_j - m)$ with $m = \max_j \ell_j$. These conditional probabilities depend on the extraction procedure. They are not self-reported confidence, a calibrated human probability, or the observed frequency of choices across images. Score summaries keep distinct prompt and extraction procedures separate. Paired probability differences require matching scoring procedures; incomplete groups retain coverage and unavailable values.

Tool contrasts use held-tool minus no-tool within each view and camera-match stratum. Layout contrasts in this study use side-by-side minus opposite; historical source directions are converted numerically before aggregation. Viewpoint agreement and color/shape sensitivity require complete groups at the declared unit. Table 1 uses the twelve-position Distance slope, eight-cell absolute Tool/Layout magnitudes, four-view disagreement, and paired Task/Order rates defined in Appendix F. All are descriptive fixed-stimulus summaries, without independent-image significance tests.

K.3 External English reference records

The external resource has 24 records from the published 2008 Table 1 (means and standard deviations), six 2023 English region summaries, and twelve 2023 English distance summaries, giving 42 primary records. Eight additional actor means are derived from rounded 2008 published inputs with equal weights and without a derived SD. They are not 50 independent experiments. The 2008 study-level English $N = 48$ is not a per-row sample count: row counts and participant denominators remain null, and SD units are percentage points. For the 2023 side-by-side R1 record, 139 *this* plus 53 *that* responses total 192, from 32 participants each contributing six responses. Trial counts and participant counts are not interchangeable.

The 2023 region and distance tables aggregate the same source responses and must remain separate analysis granularities. Labels R1–R3 in this package denote spatial regions, not three lexical demonstrative categories in English. Q1–Q4 are 2008 design groups, not answers or reachability certificates. Candidate 2023 correspondence uses only M01/M02, M06/M07 and M11/M12 at a 0.6 spatial scale; intermediate markers have null candidate mappings. These six anchors are geometric design correspondences, not established behavioral equivalence. No interpolation creates current-image gold labels.

The external summaries are not assigned as labels to individual benchmark images. The 2023 human data have no tool condition. Separate declared aggregation rules support the exploratory comparisons in Appendix E, not image-level agreement. External human data are referenced separately through their original sources; the planned public benchmark release comprises the synthetic stimuli and evaluation resources. No new human study or human recognition validation is claimed.

Exploratory comparisons require an explicit correspondence between image conditions and human conditions, normalized weights, and a specified estimator. Candidate-probability comparisons additionally require a common scoring procedure across all conditions, including those with missing predictions; incomplete cells are not silently reweighted. The analyses in Appendix E instead use generated-response frequencies and remain separate from candidate-probability scoring.

Distribution comparison uses base-2 Jensen–Shannon divergence and takes its square root within each condition. Macro distance averages those condition-level distances rather than taking the square root after averaging divergences. Absolute *this*-probability error is reported separately. Neither measure creates image-level correctness labels.

L Reproducibility and benchmark availability

The evaluation implementation has 29 recorded CPU tests using synthetic examples. They cover parsing, input separation, denominator and coverage handling, probability normalization, and scoring comparability. These software checks are distinct from geometric validation and model inference. The response analysis recomputes summaries from saved outputs; it does not rerun the earlier simulation or physical-validation studies.

Data consistency checks cover record integrity, agreement between tabular representations, and loading of individual metadata record sets. Sequential loading with a shared metadata-loader instance remains unreliable, so exhaustive interoperability is not established. These checks also do not establish complete offline re-rendering without the original simulation assets.

An anonymized benchmark package will be provided for peer review, followed by a freely accessible public release upon acceptance. The release will include the benchmark images, scene and image annotations, matched-comparison identities, evaluation prompts, evaluation code, and documentation. Version identifiers, citation information, and third-party source attributions will accompany the release. Documented annotation findings and the scope of external-human comparisons remain part of the resource description (Appendices I and E).

M Transfer from human experiments and comparison with related benchmarks

M.1 What is transferred

The human and model studies share selected experimental contrasts, but differ in the process producing responses. In the memory game, participants physically point, act on objects, and maintain an interaction history. The tested VLM receives one rendered snapshot and is instructed to answer for MIMo. Table 19 records this distinction explicitly.

Coventry et al. (2008) manipulate object-placement history and tool use in addition to distance and layout. In contrast, the 2023 cross-language experiment principally varies distance and addressee position; it uses six selected locations from the marked table, organized into three conceptual regions (Coventry et al., 2023). These are distinct studies, not one shared tool-intervention protocol.

Table 19: Human experimental design and its adaptation in ESCBench.

Element	Human paradigm	ESCBench
Distance	Marked positions and participant-relative regions	Twelve positions, 150–1,800 mm from the near edge
Tools	Participants use a 700-mm stick in the 2008 study	A depicted 520-mm rod versus bare hand
Interlocutors	Beside or opposite placement	Fixed MIMo/iCub roles in two layouts
Target	Colored two-dimensional geometric shapes	One 70-mm three-dimensional target, three shapes and four colors
Elicitation	Structured naming under memory-game cover	Explicit one-word or three-word prompts and diagnostics
Observation	Physical view and action history	One of four independently queried rendered views
Validation	Participant responses and reach measurements	Saved-state geometry plus recorded model responses

ESCBench adopts neither the original memory cover nor the history of placing each object, and it makes no role-swap or attention-manipulation comparison.

The formal selected contact ranges are M01–M02 for bare hand and M01–M05 for held tool. They do not validate correspondence to human quadrants or regions. Human summaries are not assigned as image-level labels; a separate exploratory comparison design is described in Appendix E. In particular, a published distribution normalized within a demonstrative across regions cannot be substituted for the frequency of that demonstrative conditional on a region. The former describes $P(\text{region} \mid \text{term})$; the latter describes $P(\text{term} \mid \text{region})$. Current-image participant data and reviewed correspondence rules are required for a human-agreement score.

M.2 Relation to the closest evaluations

Table 20: Evaluation scopes and the comparisons they motivate in ESCBench. The proposed increment is a matched context-by-observation analysis of speaker-conditioned demonstratives.

Resource	Documented evaluation scope	Comparison supported by ESCBench
COMFORT (Zhang et al., 2025b)	Reference-frame ambiguity, robustness, consistency, and flexibility	Response to tool and interlocutor configurations in participant-conditioned demonstrative selection.
MMSI-Bench (Yang et al., 2026)	Spatial reasoning across multiple images	Matched contextual contrasts compared across separately queried observations.
SCOPE (Kim et al., 2026)	Single-view consistency and updating; multi-view integration across egocentric and allocentric tasks	Direction and distance profile of a fixed speaker's tool/layout response across views.
Dong et al. (2026)	Vocabulary, possessives, demonstratives, human comparison, cue ablations, prompt variants, and object/speaker checks	Repeated tool and layout contrasts across distance and observation, interpreted alongside recognition and elicitation checks.
Watanabe et al. (2026)	Four-language demonstratives, three-word outputs, distance conditions, human distributions/JSD, and recognition-conditioned analysis	View-specific contextual contrasts that distinguish behaviors with similar aggregate human resemblance.

Dong et al. establish that perspectival-word use can dissociate from vocabulary and identification performance; this gap motivates our recognition checks. Watanabe et al.'s multilingual comparison supplies precedent for distribution-based and recognition-conditioned evaluation. SCOPE evaluates spatial consistency at both single-view and multi-view levels. The additional comparison here repeats the same tool or layout contrast within each observation while retaining the speaker task, then asks whether its signed distance profile persists. The reversals and aggregation examples in

Sections 5.2–5.3 show how this comparison changes the interpretation of recognition, stability, and resemblance summaries.

N HISTORICAL PHYSICAL EVIDENCE: OVERVIEW AND EXACT SCOPE

This appendix summarizes earlier physical-validation studies used during scene development. Their disk and blue-sphere tasks are distinct from the benchmark stimuli defined in Section 3.

N.1 PROBLEM AND EVIDENCE TYPES

Directional alignment, static geometric reachability, dynamic contact execution, and target visibility are distinct properties of a spatial action. A ray can intersect a target that the body cannot touch; a geometrically admissible endpoint need not admit a feasible motion or stable grasp. Conversely, an unreachable object can still be indicated successfully. We therefore treat reachability as a protocol-dependent benchmark attribute, not a prerequisite for all spatial reference. This historical module supplies physical evidence; it is separate from the new formal-image language-behavior evaluation.

The scene uses MIMo (Mattern et al., 2023) in MuJoCo (Todorov et al., 2012), with a seated observer and externally constrained bases. The nominal seat distance is 150 mm, measured from the hip-flexion anchor midpoint to the near table edge at world $x = 100$ mm. The original targets are radius-7.5-mm disks on the physical tabletop at $z = 425$ mm, with 150-mm center spacing. M02, M03, M05, and M06 lie 300, 450, 750, and 900 mm beyond the edge. Those disk trials contain no target ball. Figure 17 shows their layout and two positive witnesses; separate blue-sphere witnesses are introduced below.

For protocol P, let Q_P denote the allowed configurations, $S_P(q)$ the designated physical contact surface, and T_i the target surface (D_i for a tabletop disk). A static witness satisfies the registered surface-to-target tolerance and all geometric gates. A dynamic witness additionally executes the registered contact and hold conditions. A conditional exclusion bounds every covered surface point for every $q \in Q_P$. Finite searches without either witness or exclusion receive *unknown*; Table 21 preserves these distinctions.

N.2 DYNAMIC AND STATIC POSITIVE WITNESSES

M02 has one nominal-condition dynamic witness using the right index distal endcap. Preparation lasts 0.800 s; extension to the complete qualified hold takes 4.532 s. The final interval, 5.332–7.332 s, provides 2.000 s of actual light contact while satisfying the registered posture, hand-shape, support, and gaze conditions. Across the full recording, peak normal contact force is 0.524 N, minimum non-target table clearance is 3.822 mm, maximum actual torso tilt is 17.984°, and maximum signed actuator-control utilization is 81.657%, with no saturation. The 1-N force ceiling and 3-mm clearance floor are experimental engineering criteria, not universal human physiological thresholds. Appendix O defines the contact region and measurement windows and retains the unsmoothed actual curves. Three tuning candidates are not random trials; the archived independent replay confirms reproducibility without increasing the independent sample count. No success rate is estimated.

M05 provides a different, static geometric witness using a 520-mm rod and a newly selected grasp. The specified 5-mm-radius tip sphere lies 0.040 mm above the physical target center, passing the registered 0.1-mm static tolerance after independent reload. Native constraints, geometric grasp conditions, left-hand support, and non-target collision checks pass. The actual torso tilt is 17.879° under this search's 18° engineering ceiling. This is tolerance-based geometric contact, not observed force-bearing contact: no approach, dynamic grasp stability, hold, or natural movement has been demonstrated. Moreover, the original noncolliding visual cylinder differs from the physical shaft and tip sphere; its end rim overlaps the tabletop by 0.328 mm. The rendered geometry is retained and this limitation is disclosed.

Separately, archived 70-mm blue-sphere tasks provide static witnesses at bare-hand M03 and rod-assisted M06. Their positive surface gaps are 0.0200 and 0.0552 mm, within the respective 0.05- and 0.10-mm tolerances; neither establishes force-bearing contact, stable grasping, or dynamic execution. These configurations use substantial lean and turning, without the old 18° ceiling; the tool protocol also permits axial regrasping. Geometry, contact height, posture and grasp policy are not

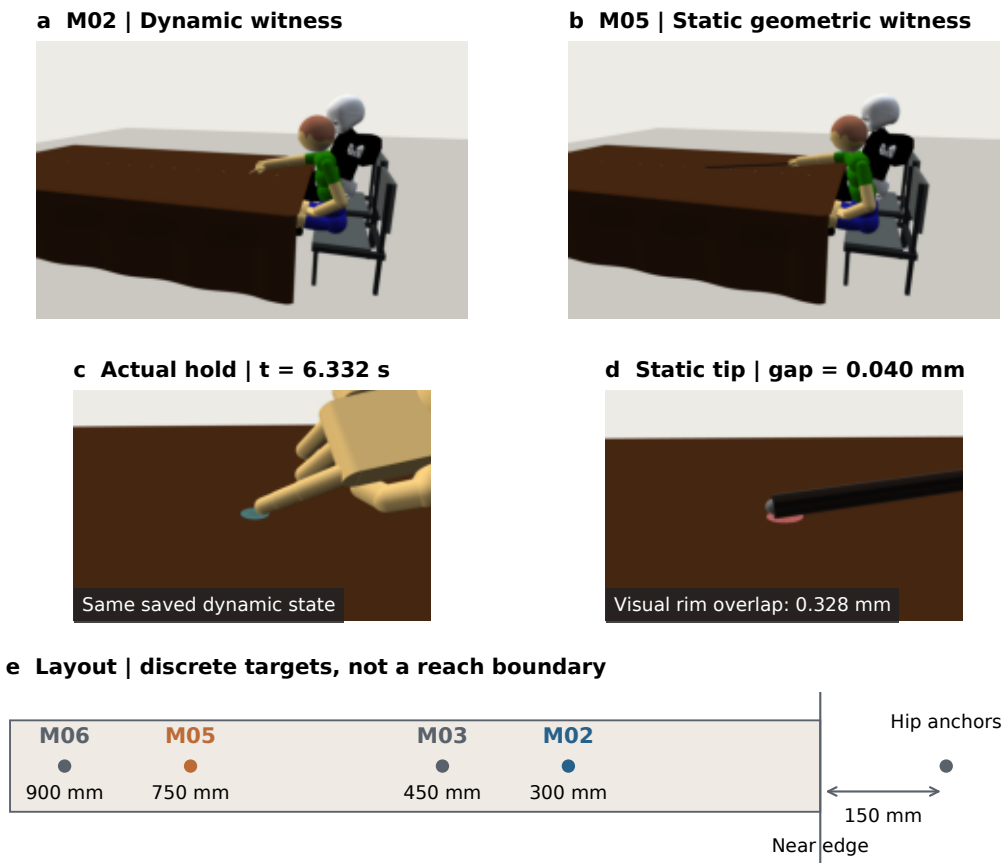


Figure 17: Two evidence types in the same fixed world view. (a) M02 actual dynamic hold at $t = 6.332$ s from the saved trajectory. (b) M05 final saved static pose; no movement or contact force was verified. (c,d) Local views of the same states. M05's designated physical tip has a positive 0.040-mm gap within its static tolerance; the retained noncolliding visual rod rim overlaps the table by 0.328 mm. (e) Discrete target layout, with distances from the near edge; omitted intervening targets do not imply a boundary. The 150-mm seat dimension uses hip-flexion anchors. The diagram does not represent a continuous reachable region.

controlled factors, so success cannot be attributed solely to sphere size. The disk exclusions and unknown labels remain unchanged. These witnesses establish neither the farthest reachable position nor unreachability of later positions. Appendix S records the exact states, constraints and images.

N.3 Conditional exclusion and expanded-domain unknowns for disks

The earlier M03 result is a *computer-assisted conservative exclusion certificate*. Its Q_{18} retains native joint and tendon constraints with 10^{-5}-rad tolerances, limits actual upper-body tilt to $18°$, and permits pelvis translation within 2 mm of a specified anchor. It additionally assumes pelvis rotation within $0.25°$ of nominal identity. This angular envelope is a certificate assumption, not an automatic acceptance guarantee of the original dynamics protocol. The $18°$ ceiling is an engineering posture constraint, not a mechanical limit.

The final tolerance-expanded certificate encloses the whole designated distal capsule, including its permitted contact surface, and establishes

$$\forall q \in Q_{18},\ \forall s \in S_P(q):\ s_x < 540\ \text{mm}, \qquad \forall d \in D_3:\ d_x \geq 542.5\ \text{mm}. \tag{3}$$

Thus the entire surface and disk are conservatively separated by at least 2.5 mm. This is neither an exact maximum reach nor a failed dynamics trial, and it excludes no other body part. Appendix P

Table 21: Historical protocols only. Protocol-indexed status. Unmarked targets are radius-7.5-mm tabletop disks; ball rows use 70-mm blue spheres. Edge distance refers to the target center. Search gaps describe saved poses, not global minima. "Excluded" applies only to the stated certificate domain; unknown is not an unreachable negative label. Force and hold are unevaluated (—) for static and certificate evidence.

Target	Edge (mm)	Main protocol	Evidence	Key quantity	Status
M02	300	Bare hand; 150-mm seat; tilt $\leq 18^\circ$	Actual dynamics	0.524 N peak; 2.000 s hold	Dynamic positive
M03	450	Bare hand; Q_{18}; extra pelvis envelope	Entire-surface certificate	≥ 2.5 mm separation	Conditional exclusion
M03	450	Expanded angles; original / forward 10 mm	Static search	27.574 / 28.005 mm gap	Unknown
M05	750	New rod grasp; tilt $\leq 18^\circ$	Static geometry	0.040 mm gap; tolerance 0.1 mm	Static positive
M06	900	Expanded angles; fixed M05 grasp and handshape	Static search	18.360 mm full-surface gap	Unknown
M03 ball	450	Bare hand; native angles; retained contact alternative	Static geometry	0.0200 mm gap; tolerance 0.05 mm	Static positive
M06 ball	900	Rod; native angles; axial regrasping permitted	Static geometry	0.0552 mm gap; tolerance 0.10 mm	Static positive

retains the domain coverage and numerical audit, including checks of ordinary floating-point length constants; the complete pipeline is not formally verified.

Removing artificial angular caps changes the question. Expanded-domain bare-hand M03 searches yield best independently checked gaps of 27.574 mm in the original layout and 28.005 mm after both seated setups move forward 10 mm. These saved-pose distances are not global separation lower bounds. The original-layout pose has 27.861° torso tilt, substantial lateral turning, and only about 5 nm chest–tablecloth clearance: it is a boundary-sensitive diagnostic, not a robust or natural endpoint. The M03 disk remains unknown in this expanded domain.

For the M06 disk, the enlarged search fixes M05's new hand–rod transform and all 22 grasping-finger angles while removing artificial torso, wrist, and other posture caps. Its best retained noncontact pose has a full tip-sphere-to-disk distance of 18.360 mm. The separate 22.805-mm horizontal gap measures the lowest tip point's projection and must not replace that surface distance. A conservative world-coordinate enclosure gives $X \leq 1.114899$ m; this is not an attained reach. Its overlap with M06 proves neither reachability nor unreachability. The M06 disk therefore also remains unknown, specifically for the fixed grasp.

N.4 Benchmark interpretation and scope

Labels should carry their protocol and evidence type: dynamic positive, static positive, conditional exclusion, or unknown. M02 and M05 use different contact objects and conditions, so they do not quantify an exact tool benefit or reach multiplier. The formal collection now supplies matched 70-mm-object tool pairs, but not a bare-hand exclusion for the same historical M05 disk; that strict comparison remains future work. External support precludes claims of autonomous seated balance or hardware executability. M02's visual axes face the target while its forearm/hand occludes the contact; its hand-back deviation remains about 45.3°. M05's target is tool-occluded. Alignment therefore does not establish visibility.

The layout is inspired by Coventry et al. (2008): their no-tool mean reach lay between dots 2–3, and tool-group reach between dots 4–6. These regional Q1/Q2 descriptions do not guarantee every participant's access to every dot. Their 25-cm spacing and 70-cm rod differ from our 15-cm spacing and 52-cm rod, as do morphology and constraints. These observations motivate the design without establishing agreement with the human regional arrangement or a quantitative boundary or language-behavior replication. It does not establish a complete workspace, disturbance robustness, or improved communication success.

O M02 DYNAMIC ACCEPTANCE

O.1 ACTUAL CONTACT PROTOCOL AND MEASUREMENT WINDOWS

The experiment registers engineering acceptance conditions for a fixed, supported seated configuration. These are simulation-specific adaptations, not a replication of a human experiment or a claim that human movement obeys the same thresholds. The 150-mm distance uses the mean of the two actual hip-flexion joint anchors and the near tabletop edge at $x = 100$ mm. MIMo's initial anchor midpoint has $x = -50$ mm. Its pelvis body origin differs from this measurement reference. The visual marker center is 0.15 mm above the physical tabletop; the contact target is its circular footprint on the physical top face, not the visual cylinder.

Active contact and region. An active contact has nonpositive signed solver distance, a valid constraint address, no exclusion flag, and positive normal force. A valid target contact must involve the right index finger's distal collision capsule and the physical tabletop collision surface. Its solver contact point, projected into the tabletop plane, must lie within the original 7.5-mm radius. The distal capsule axis must point toward the tabletop. With capsule center c, axis u, and half-length h, its permitted endcap test is $(p - c)^\top u \geq h - 0.5$ mm. The top-face normal cosine must be at least 0.99, and the top-face position tolerance is 0.5 mm. These tests exclude the shaft, proximal cap, outside-disk contact, and table sides or underside.

Normal force for contact qualification is the sum over valid target contacts. The upper force check also sums all active designated distal-table contacts, so invalid distal contacts cannot evade the 1-N cap. These sums coincide in the selected recording. Required force is 0.02–1 N and maximum target penetration is 0.5 mm. The maximum measured target penetration is 0.013505 mm. Actual closing speed is limited to 0.03 m/s in the registered approach zone (phase at least 3, normal gap at most 10 mm, planar center distance at most 17.5 mm) and at every contact-entry sample and its predecessor. The archived independent maximum is 0.025649 m/s. The wider approach zone does not enlarge the contact disk.

Qualified hold. The final interval must satisfy valid touch, the observer's target-gaze and permitted-occluder conditions, MIMo's bilateral target-axis error bound of 3°, the actual torso condition, and the terminal hand-shape/support check at every recorded state. The observer's final gaze error must be at most 1°; this protocol permits the contacting index to be the first occluder. MIMo's own occlusion is reported separately and is not converted into a claim of visibility. The terminal check requires index PIP/DIP magnitudes at most 12°, hand-back deviation at most 50°, and at least 95% qualified samples for each required hand and seat/foot support group over the recent 2 s. The non-reaching hands rest on their same-side thighs. MIMo starts upright within 3°; its lean branch requires at least 3° actual forward lean and maintains the 18° total-tilt ceiling. Whole-run safety and independent final-window checks must also pass.

The online complete-condition clock starts at 5.332 s and runs continuously to 7.332 s. The independent final-window audit tests the same final 2-s suffix, including both endpoints. Near-touch alone is first valid at 5.219 s and becomes continuously valid from 5.244 s; neither earlier time replaces the complete-condition start. The 0.800-s preparation plus the 4.532-s extension-to-hold interval plus the 2.000-s hold equals the 7.332-s record. Numerical endpoint tolerance is 10^{-9} s.

Clearance, tilt, and utilization. The full-run extrema are 0.524075234 N contact force, 3.821576818 mm non-target table clearance, 17.983731640° actual torso tilt, and 81.656954007% signed control utilization. They cover all states from 0 to 7.332 s. The hold force ranges from 0.086390 to 0.150863 N. Figure 18 retains every saved sample. The distance audit evaluates

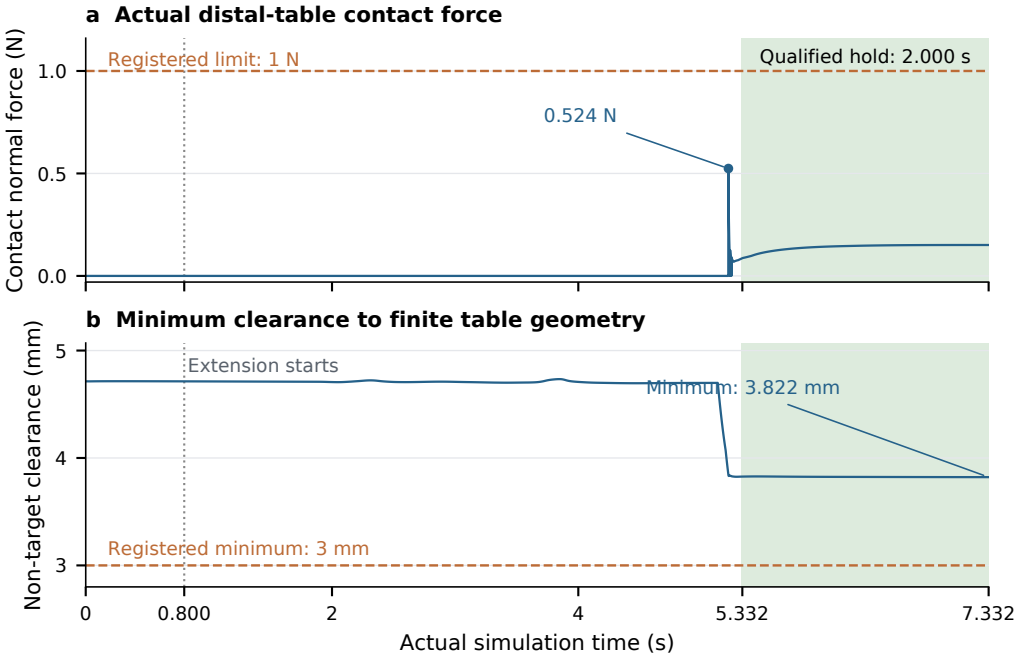


Figure 18: Original actual M02 records at 1-ms spacing (7,333 states). Contact normal force sums active designated distal-table contacts; the qualified-contact sum is identical in this record. Clearance is the saved conservative minimum over finite non-target table geometry. There is no smoothing, interpolation, or reference-trajectory substitution. Both panels share actual time; the shaded interval is the final qualified hold, 5.332–7.332 s, containing 2,001 states including endpoints. Thresholds are this protocol's engineering criteria.

robot/chair geometries against finite tabletop and leg geometries, with analytic or conservative primitive/mesh distance bounds and explicit treatment of uncertain pairs. The plotted value is the saved per-state minimum conservative non-target distance; it is not an estimate of distance to an infinite table plane. Only the designated target endcap is exempted during the registered approach/contact conditions. A broad-phase overlap or unresolved bound below 3 mm cannot yield a pass. The independent audit confirms the physical-table minimum of 3.821576817933 mm, at 7.332 s, between the index middle segment and the physical slab; the coincident visible slab gives the same value.

For actual upper-body world superior axis $v = R_{\text{upper}}(0,0,1)^\top$, total tilt is $\operatorname{atan2}(\sqrt{v_x^2 + v_y^2}, v_z)$. It is not the sum of waist joint angles. Signed control utilization for finite limits $l_j < 0 < u_j$ is

$$\rho = \max_{t,j} \begin{cases} c_j(t)/u_j, & c_j(t) \geq 0, \\ c_j(t)/l_j, & c_j(t) < 0. \end{cases} \tag{4}$$

The maximum is 0.816569540071 at 3.590 s for the MIMo left hip actuator. This statistic describes control-limit use, not energy efficiency or an independently measured torque robustness margin. Actual actuator force and gear-transformed joint torque were also checked against the frozen limits; saturation count is zero.

O.2 Execution and supporting measurements

The supported bases retain active world welds; the upper-body weld is inactive. No live base reset or externally applied force is used after initialization, but weld constraint forces are present. For example, the archived mean vertical MIMo weld force during hold is 160.209 N. Zero applied-force fields must therefore not be interpreted as unsupported balance. The left-hand controller tracks its initial pose relative to the moving thigh using finite arm motors. Registered downward motor-task preloads are 1.5 N for MIMo resting hands and 0.5 N for observer hands; observer foot preload is

Table 22: Execution details and supporting actual measurements. Controller gains are the registered values in their implementation units. They are engineering settings, not fitted human parameters.

Item	Setting or actual result
Simulation environment	MuJoCo 3.3.0 with Python-based numerical control
Time step; gravity	1 ms; $(0, 0, -9.81)$ m/s^2
Integrator; solver	implicitfast; Newton, 100 iterations, tolerance 10^{-10}
Friction solver	Elliptic cone; impedance ratio 100; no-slip iterations 0
Arm / finger PD gains (K_p, K_d)	$(35, 1.6)$ / $(0.25, 0.03)$
Torso PD gains (K_p, K_d)	$(250, 20)$
Seated leg / hip-pitch PD gains	$(35, 2)$ / $(350, 10)$
Contact force target; integral gain	0.15 N; 6 s^{-1}
Downward command rate; cap	0.5 N/s; 1 N, through finite motors
MIMo left-thigh hand support	95.827% whole run; 95.316% after preparation
Maximum left support penetration	0.079264 mm (limit 1 mm)
Maximum fixed-support translation	0.055927 mm (limit 2 mm)
MIMo pelvis orientation	0.087691° maximum from world identity; 0.016290° from first saved state
MIMo hold visual-axis error	0.0027–0.0276° across both eyes
Observer maximum hold target error	0.7576°
Hand-back hold deviation	45.321–45.350°

1 N and MIMo foot preload is zero. Compensation and preloads share the existing actuator limits. The MIMo head and each eye retain three native joints. The controller uses actual torso state when updating gaze, with head reference speed/acceleration limits of 60°/s and 240°/s^2, and eye limits of 120°/s and 600°/s^2.

An independent replay previously reproduced the 7,332 integration steps. The reported curves use the original trajectory, and images are rendered from saved states without additional time integration. No timing ablation was performed.

MIMo's target axes are within their registered angular bounds during the hold, but all 2,001 hold samples remain self-occluded: the first ray intersections are the right forearm or hand. Alignment is therefore reported separately from visibility. The approximately 45.3° hand-back deviation passes the registered 50° bound but leaves room for improved naturalness.

P Conditional M03 exclusion certificate

This certificate applies only to the stated $\leq 18^\circ$ torso domain, denoted Q_{18} in Section N. It does not apply to the expanded-domain searches in Appendix Q. The designated surface $S_P(q)$ is the right index contact endcap; the bounding construction below covers the entire distal capsule.

P.1 Configuration domain and entire-surface enclosure

The geometric set Q_{18} fixes the compiled morphology and table targets. The nominal pelvis body origin is

$$P_0 = (-54.9912617863, 130, 267.7240391293)\ \text{mm}, \tag{5}$$

with identity nominal orientation, translation satisfying $\|P - P_0\|_2 \leq 2$ mm, and rotation distance to identity at most $\gamma = 0.25^\circ$. The position envelope is stated relative to this certificate anchor; it does not assert equivalence to every possible dynamic initial condition. The certificate's angular envelope is an additional assumption, distinct from the protocol's 2-mm drift gate. It does not cover arbitrary rotation following a soft-weld failure.

Write the first waist angles as (a, r, b) and the chest angles as (c, y) in the compiled order. Their native ranges in degrees are $[-14, 14]$, $[-18, 18]$, $[-17, 30.5]$, $[-9, 9]$, and $[-18, 18]$. The final certificate expands every corresponding compiled binary64 radian interval by $\tau = 10^{-5}$ rad, with outward rounding. Joint limits for the second waist and all arm/finger rotations are relaxed in the bounding computation; this includes the original allowed ranges and their tolerance. Other motion, collision, support, and naturalness restrictions can only shrink the original task set.

Native waist tendons impose

$$|a - \tfrac{1}{2}a_2| \le L_a, \qquad |r - r_2| \le L_r, \qquad |b - b_2| \le L_b, \tag{6}$$

where each accepted half-range is bounded outward by $L_i = 0.010010000000000002$ rad, including the compiled native endpoints and tolerance. Thus, $a_2 = 2a + \delta_a$, $r_2 = r + \delta_r$, $b_2 = b + \delta_b$, with $|\delta_a| \le 2L_a$, $|\delta_r| \le L_r$, and $|\delta_b| \le L_b$. The rotation-distance discrepancy from the nominal second waist is bounded by $\epsilon = 2L_a + L_r + L_b \simeq 0.04004$ rad.

Let $R_1 = R_x(a)R_z(r)R_y(b)$ and $R_2^* = R_x(2a)R_z(r)R_y(b)$. Using fixed offsets in mm,

$$p_1 = (2, 0, 25.585648837), \qquad v_{12} = (-2, 0, 102.342595346), \tag{7}$$

$$v_{2c} = (0, 0, 157.963571078), \qquad s = (0, -104.602469724, 0), \tag{8}$$

the nominal shoulder position is

$$P_s^* = P_0 + p_1 + R_1 v_{12} + R_1 R_2^* \big(v_{2c} + R_x(c)R_z(y)s\big). \tag{9}$$

The compiled rotation order, pivot positions, and native tendon coefficients underpin this expression. The first waist angle occurs once in R_1 and twice in R_2^*; all three occurrences are included in each interval calculation.

For the designated capsule, a triangle-inequality enclosure adds shoulder-to-elbow, elbow-to-first-wrist-pivot, the two offsets between the three noncoincident wrist pivots, third-wrist-pivot-to-index-knuckle, and the two remaining index links. Their lengths are respectively 164.830716, 145.465777, 4.360183, 4.360183, 49.368135, 26.028868, and 13.014434 mm. A further 13.014434-mm envelope contains the entire distal capsule about its joint: center-offset norm plus capsule half-length plus radius. The resulting shoulder-centered radius is $r_S = 420.442729735$ mm. This construction includes the cylinder and both spherical caps, and therefore contains all permitted contact endcap points. It is not a fingertip-center bound.

Native tendon uncertainty contributes

$$\Delta_n = 2\sin(\epsilon/2)(\|v_{2c}\| + \|s\|) = 10.512442009 \text{ mm}. \tag{10}$$

Pelvis rotation contributes

$$\Delta_p = 2\sin(\gamma/2)(\|p_1\| + \|v_{12}\| + \|v_{2c}\| + \|s\|) = 1.704277211 \text{ mm}. \tag{11}$$

It need not rotate the arm enclosure separately because the shoulder-centered sphere is rotation invariant. Every covered surface coordinate obeys

$$x \le (P_s^*)_x + r_S + \Delta_n + \Delta_p + 2 \text{ mm} + 0.010 \text{ mm}. \tag{12}$$

The final term is the numerical allowance. Actual upper-body tilt at most 18° implies nominal upper-body tilt at most $18^\circ + \epsilon + \gamma = 20.544123012^\circ$. A box is tilt-infeasible only when its upper bound on $(R_1 R_2^*)_{33}$ is below the lower bound on the cosine of this enlarged limit. Chest rotation affects shoulder position but not the defined upper-body tilt axis.

P.2 Coverage, final tolerance extension, and numerical audit

The original robust-pelvis certificate partitions the five-angle root box by bisection into 3,071 leaves. To cover joint tolerance, every original leaf is expanded in all five coordinates, including leaves previously pruned by tilt. For any point in the enlarged root box, coordinatewise clamping into the original root yields a point in an original leaf; the unclamped point is within τ per coordinate and is therefore contained in that leaf's expansion. Expanded leaves may overlap without leaving uncovered points. Each expanded parent is fully covered by its recorded bisection descendants.

The tolerance-expanded certificate revisits 1,178 old tilt-pruned leaves; 37 no longer satisfy the enlarged-domain pruning condition. Reclassification and 48 additional bisections (96 child nodes) close all 3,119 leaves with zero unresolved boxes: 1,187 tilt exclusions and 1,932 spatial exclusions. The largest spatial upper bound is 539.998699601413 mm, strictly below 540 mm. Exact-rational checking of the compiled target center-minus-radius confirms the conservative 2.5-mm separation.

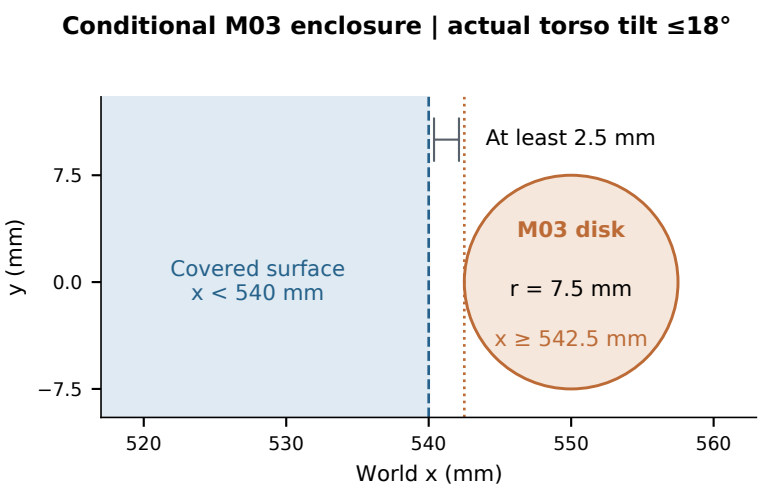


Figure 19: Conservative coordinate separation for the $\leq 18^\circ$ certificate domain, including the additional 0.25° pelvis-rotation envelope. The shaded half-plane is an outer enclosure, not an exact workspace. Every covered surface point lies below $x = 540$ mm; the entire target disk lies at or above 542.5 mm. No M03 failed posture is inferred from this schematic.

What the arithmetic checks establish. The generator uses binary64 interval operations with 10^{-13} absolute padding and adjacent-float outward rounding, plus the final 10-μm spatial allowance. An archived independent 50-decimal mpmath interval replay avoids the generator's interval implementation. The tolerance-expanded certificate uses interval trigonometry and outward-rounded compiled joint/tendon endpoints. Its exact-rational boundary audit contains 16 passing checks, including expanded-domain coverage and source/certificate identities. High arithmetic precision alone does not make the complete pipeline formally verified.

In particular, the original generator computes some link-length constants with ordinary NumPy norms; the archived replay treats their serialized values as inputs. A separate geometric check reconstructs all seven link vectors, capsule envelope, and shoulder offsets directly from the frozen compiled primitive coordinates, using outward interval subtraction, norms, and summation. Relative to that compiled geometry, the serialized arm-radius sum does not underestimate the reconstructed radius. Propagating coordinate-rounding discrepancies through the shoulder and perturbation bounds yields a total possible constant understatement below 10^{-12} m (computed upper bound 1.422×10^{-17} m), which is covered by the existing 10^{-5}-m allowance. An independently written 60-decimal replay also reconstructs the root and expanded-parent coverage and verifies every final leaf without further refinement.

The certificate trusts the compiled coordinates, nominal pelvis anchor, rigid-kinematic interpretation, and interval library. The executable audits use full serialized precision. Compiler semantics and morphology fidelity are not formally verified. An archived equivalence check confirms that the frozen no-ball model preserves the certificate's relevant skeletal, target, and constraint arrays. The exclusion remains limited to the stated configuration and contact surface.

The reported result uses the completed interval and geometric-constant checks; the full certificate search is not repeated as part of the image evaluation. This is a computer-assisted conservative exclusion certificate, conditional on the stated model and domain assumptions. The bound is not a measurement of maximum reach, and it says nothing about another seat, grasp, morphology, larger torso tilt, or another body surface.

Q Static searches and retained limitations

This appendix retains the original tabletop-disk searches. The blue-sphere witnesses have separate target geometry and protocols (Appendix S).

Q.1 EXPANDED-DOMAIN BARE-HAND M03

The supplemental static search removes the artificial torso, chest, wrist, and hand-posture caps and local joint boxes, while retaining compiled native joint ranges, limit tendons, scene geometry, and seated support requirements. Naturalness is a soft preference. Right fingers and the resting hands can change; this is not the dynamic witness's fixed hand-shape protocol. The target is the original M03 disk, centered at $(0.550, 0, 0.425)$ m. The objective is shortest distance from the actual forward distal hemisphere to the entire closed disk. Its static contact tolerance is 0.05 mm, with at most 10^{-8} m numerical penetration. The stored nearest surface/disk points are distance endpoints, not actual contacts.

The registered budget comprises 12 local calls, including support initialization and observer repair. Initial states include a previously retained native-range candidate, two alternative legal lap-support branches, and shifted layouts with repaired observer support. The shared-forward range is 0–60 mm; both seated setups move together relative to the fixed table. Two calls terminate early and still count: one observer-only solve conflicts with a fixed MIMo cloth constraint, and one combined donor-arm/torso seed creates 35.54 mm thigh penetration. Subsequent whole-pose searches and final checks cover the complete retained scene. Existing stored iterates are screened without additional optimization.

The selected original and 10-mm-forward results have gaps of 27.574270 and 28.004876 mm. Their measured MIMo/observer seat distances are 150.002/150.180 mm and 140.002/140.180 mm, respectively. The small offsets reflect the anchor-based measurement, not a changed definition of nominal seating. Table 23 records their posture and environmental margins. Independently checked 20- and 40-mm-shift candidates have gaps of 28.111 and 36.891 mm. A 60-mm observer repair passes its local checks, but whole-scene reaching candidates retain collision/clearance failures and are not adopted. Neither the nonmonotonic gaps nor similar results from two lap branches establish a global optimum.

All palm, thumb, finger, seated, and foot support geometries of both robots are checked. Support penetration may not exceed 1 mm and the registered proximity bound is 0.2 mm. The non-target physical-table clearance floor is 3 mm; abnormal native non-support penetration may not exceed 0.2 mm. That table-specific clearance does not imply 3-mm separation for every self-collision pair. A 2-μm planning reserve is recorded separately; final acceptance uses the original thresholds. The retained three-sided cloth is checked, and unresolved top-sheet capsule pairs receive a triangle-level completion.

For the original layout, that completion reports a minimum chest–cloth gap of 4.836×10^{-9} m, or 4.836 nm, passing its 10^{-9}-m numerical gate. The torso tilt is 27.861° and chest yaw 59.724°. Such minute clearance and pronounced turning should not be read as a physically robust or natural posture. The forward layout has a 1.994-μm chest–cloth gap and reaches 0.198 mm tolerated non-support penetration. Both are static diagnostics without forces, motion paths, or balance verification.

Table 23: Saved noncontact configurations. Positive target gaps are local measurements, not global lower bounds. The M06 value uses the entire tip sphere. These rows all retain unknown reachability status in their expanded domains.

Quantity	M03 original	M03 forward 10 mm	M06
Surface–disk gap (mm)	27.574	28.005	18.360
Actual torso tilt (°)	27.861	24.812	34.594
Chest yaw (°)	59.724	58.740	60.340
Non-target table clearance (mm)	3.000	3.002	3.061
Chest–cloth top gap (μm)	0.004836	1.994	—
Dynamic force / hold	—	—	—

Q.2 M05 STATIC POSITIVE AND VISUAL MISMATCH

The final M05 pose is the preserved first physical pass from the refined 35-degree grasp branch. It is not replaced by a later image or a changed handshape. Search records contain eight grasp

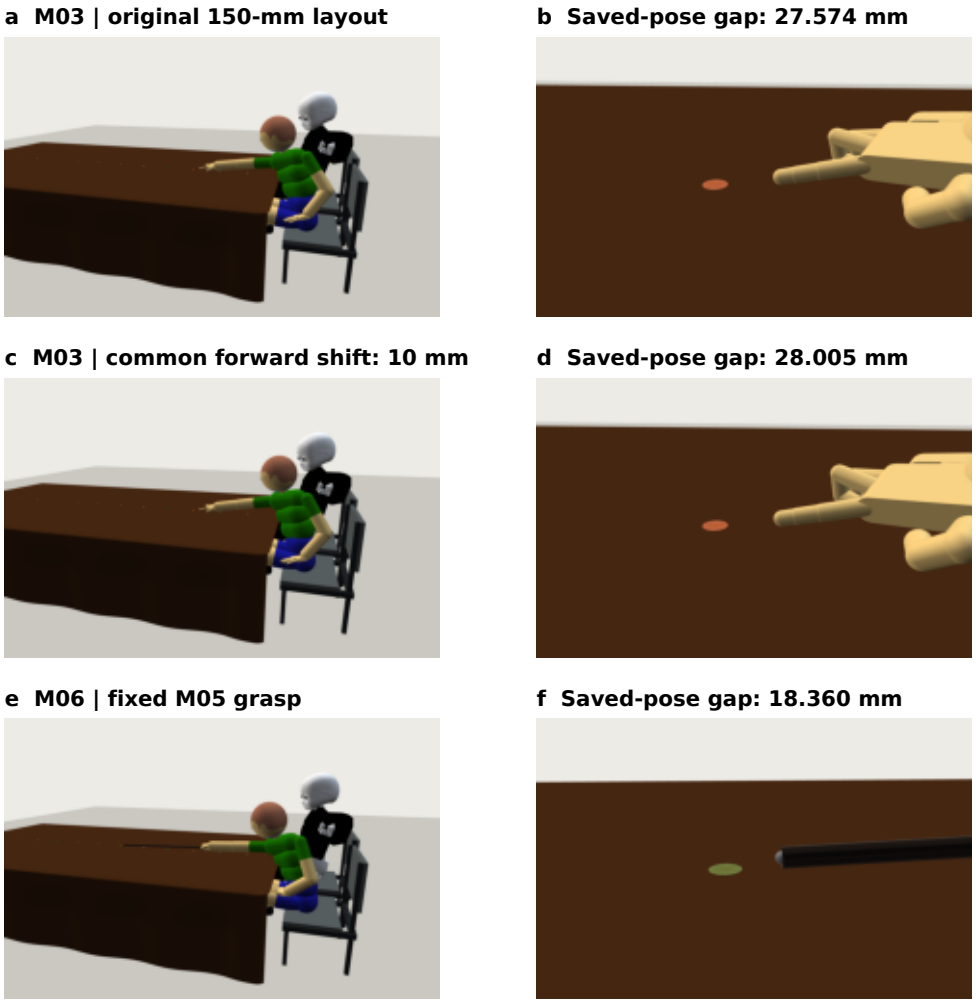


Figure 20: Exact saved noncontact states, rendered without time integration. Left: identical fixed world camera. Right: target closeups of the same states. (a,b) M03 at nominal 150-mm seating. (c,d) Both seated setups shifted forward 10 mm. (e,f) M06 with M05's fixed grasp and handshape. The displayed gaps are reconstructed from the designated physical surface and the whole disk; they are not inferred from visible marker thickness. M03's original-layout chest–cloth gap is only about 5 nm. M06 retains a 0.167-mm overlap of the noncolliding visual rod rim with the table. No pose is claimed to be a failed executed motion, natural endpoint, or boundary of reachability.

optimizations and 13 static pose solves: four main searches followed by nine visual-rim adjustment attempts. Seeds are a previous natural-posture reference and an admitted M04 static reference; a historical dynamic failure frame is not used. Tested axial hand-origin stations are 40, 28, 16, and 4 mm from the tail, with adjacent-solution warm starts; direction comparisons include 35°, 42°, 45°, and 50°. An erroneous initial 45-degree warm start mixed an old axial center-of-mass quantity into a transverse coordinate; the corrected and failed records are retained. No failed visual refinement replaces the physical-pass pose.

The selected hand origin projects 4.000 mm from the tail, the most rearward tested passing station, not a proven optimum. Five lateral near-contact points involve the thumb, index, and middle fingers, including proximal segments. They span 9.467–31.516 mm from the tail; the largest uncovered cross-sectional arc is 146.536°. This is not five-finger contact. Acceptance requires a thumb point, two non-thumb fingers, a proximal/nonterminal segment, gaps at most 0.2 mm, palm gap at most 1 mm, hand–rod penetration at most 1 mm, and an uncovered arc at most 180° with points on the rod's lateral surface. These geometric criteria do not establish frictional load capacity.

The independently reloaded pose has 0.023925-mm maximum hand–rod penetration and 0.602244-mm palm gap. All 21 left-hand geometries are checked against the thigh: maximum penetration is 0.700 mm, palm gap 0.100 mm, and left-arm/hand–trunk clearance 12.248 mm. Native joint violation is zero; the largest tendon residual is 8.7×10^{-18} rad. The physical shaft–table gap is 0.054677 mm and the non-target robot/chair–table bound is at least 4.713268 mm.

The unchanged 520-mm visual cylinder is noncolliding and has a different front surface from the physical shaft plus specified sphere. Its front-rim overlap is 0.328144 mm relative to the physical top, or 0.353144 mm relative to the cloth top. The pose satisfies the physical criteria but retains a visible rod–table intersection. No mesh has been hidden, repaired, or made supportive in the paper renders. The physical tabletop is at $z = 0.425$ m, the cloth top about 25 μm above it, and the visible marker center at $z = 0.42515$ m; physical contact is not judged from the visible marker. MIMo's two actual eye-axis errors are 0.000671° and 0.000064°. The observer's mean pupil axis aligns, but its nonvergent individual eyes each differ by about 2.064°. The checked axes first hit the rod, so the target is occluded.

Q.3 M06 WITH NATIVE POSTURE RANGES AND A FIXED GRASP

The M06 exploration keeps the new M05 hand–rod transform and 22 right-finger angles fixed, while allowing the torso, arms, wrists, and left fingers to move through native ranges. Later refinement includes the three head joints. It removes the 18° ceiling and all additional chest-yaw, lateral-tilt, hand-back, wrist-bend, and elbow gates; there is no replacement 30° cap. The native tendons and couplings remain. Numerical optimization uses a 10^{-7}-rad inward offset; independent acceptance uses native endpoints with the original 10^{-5}-rad tolerance. Gaze and posture angles are descriptive in the final native-domain audit. Naturalness has low or zero objective weight, not an acceptance gate.

The search records 16 terminal configurations, not 16 independent random trials or a global search budget. They include eight static continuation targets at world $x = 0.870, 0.890, 0.910, 0.930, 0.950, 0.970, 0.985, 1.000$ m, constraint refinement, direct extension maximization, and varied waist/shoulder/wrist starts. A further above-table initialization addresses a local solution below the table. These continuation points are numerical seeds without time parameterization, not an executed trajectory. One initialization with an inadmissible gaze target terminates before a pose is saved and is excluded from the retained configurations.

The retained best state is the $x = 0.970$ continuation result. Its only failed final acceptance item is the M06 disk. With tip center c, radius $r = 5$ mm, and disk D_6, its full-sphere distance is $\mathrm{dist}(c, D_6) - r = 18.360396864$ mm. The bottom-point projection has a separate horizontal disk gap of 22.805241354 mm, while that bottom point is only 0.062520 mm above the table. The nearest sphere/disk pair is approximately $(0.974576, 0.000025, 0.428979)$ and $(0.992500, 0.000007, 0.425000)$ m. It is not a contact pair.

Torso tilt is 34.594° and chest yaw 60.340°. The closest chest–table pair has 3.061224-mm clearance, only 0.061224 mm above its floor. Further local extension is obstructed by table clearance and nearby native waist/chest/wrist limits; for example, the failed $x = 1.000$ target pose has 3.148-mm

torso–table penetration. Some high-lean starts fall below the table. These explain observed failures without excluding other admissible configurations. The retained pose's largest tendon excess, 1.724×10^{-7} rad, is within the numerical tolerance. Its original visual-rim overlap is 0.167209 mm; the physical shaft remains 0.068058 mm clear.

The new cheap outer enclosure relaxes upper-torso rotations to complete rotation balls, retains first-waist native range and the fixed last-wrist-to-tip geometry, and uses no 18° pruning. It reports world $X \leq 1.114898879$ m for the entire tip sphere, whereas the conservative M06 disk lower X is 0.992499999 m. Thus it cannot exclude M06. This triangle-inequality enclosure uses compiled constants, ordinary floating-point calculations, and a 10-μm spatial reserve; it has not received the full interval audit of the older M03 certificate. Its overlap is only an inconclusive outer-bound result, never an attained reach or a reachability proof. Old-grasp tool certificates are inapplicable and are not reused.

R PROTOCOL COMPARISON AND REUSE

All historical disk conditions in this appendix use frozen compiled models in the MuJoCo 3.3.0 environment, retain the physical table and three-sided cloth, and define disks on the physical top face. Scene changes in the 10-mm-forward variant are an explicit layout intervention, not a perturbation trial.

The non-target robot/chair–table floor is 3 mm, with the registered target exception only where applicable. Static support penetration is limited to 1 mm and abnormal non-support native penetration to 0.2 mm; these are distinct from the table floor. M05/M06 physical shaft–environment separation and geometric grasp checks are additional requirements. Complete retained cloth checks and reported visual-surface mismatches should accompany the labels. Joint/tendon numerical acceptance uses 10^{-5} rad. All these values are engineering conventions for this simulation, not inferred human tolerances.

Directional alignment and visibility are also separate fields. M02 reports actual eye axes with self-occlusion. M05 reports tool occlusion despite aligned axes. M06's axes are aimed toward M06; the right MIMo eye first intersects the tool, while the other checked axes first meet the target marker. A marker-ray hit is not evidence of an unobstructed view of every physical contact point. No condition supplies human comprehension, demonstrative-choice, or communication-success labels.

The historical design comparison in Appendix N is restricted to the original study's reported reaching regions and apparatus dimensions (Coventry et al., 2008). It supplies no individual-level label transfer. The simulation's posture, support, tolerance, and grasp conventions are its own adaptations. A same-M05-target bare-hand exclusion plus tool witness, dynamic execution of the new grasp, improved visual clearance/naturalness, and disturbance tests remain separate future studies; none is silently assumed complete here.

Table 24: Original tabletop-disk protocols. Protocol differences that must accompany benchmark labels. Native ranges and limit tendons are retained throughout. The certificate uses conservative relaxations that enclose its allowed configurations. A dash means a quantity is not evaluated.

Property	Bare hand M02 / M03 certificate	Expanded M03	Tool M05 / expanded M06
Scene configuration	Original seated apparatus; equivalent skeletal geometry for the certificate	Original layout and a shared 10-mm forward shift	Same rod apparatus for both targets
Nominal seat	150 mm	150 mm / 140 mm after common shift	150 mm
Contact object	Right distal index endcap; certificate encloses whole capsule	Right forward distal hemisphere	Specified 5-mm-radius rod-tip sphere
Handshape / grasp	Dynamic terminal PIP/DIP $\leq 12^\circ$; certificate relaxes arm/fingers	Fingers and lap support may change	New grasp chosen for M05; transform and 22 right-finger angles fixed for M06
Artificial angles	Actual torso $\leq 18^\circ$; dynamic terminal hand-back $\leq 50^\circ$	Removed; naturalness soft	M05: torso $\leq 18^\circ$, yaw/back $\leq 20^\circ$, side $\leq 5^\circ$, wrist bends $\leq 35^\circ$; M06: removed
Support	World welds; hand/seat/foot checks; certificate adds 2-mm/0.25° pelvis envelope	Fixed seated roots; full hand/seat/foot geometry	Fixed roots; left palm and all 21 hand geometries; no dynamic grasp test
Target tolerance	Active force 0.02–1 N; penetration ≤ 0.5 mm; certificate separation ≥ 2.5 mm	Static gap ≤ 0.05 mm; no pass found	Static gap ≤ 0.1 mm; M05 pass, M06 not found
Recorded force / hold	M02: 0.524 N peak / 2.000 s; certificate: —	—	—
Evidence type	Dynamic witness / conditional exclusion	Saved-pose searches; unknown	Static witness / saved-pose search; unknown

S HISTORICAL STATIC WITNESSES FOR THE 70-MM BLUE-SPHERE TASK

S.1 TARGET GEOMETRY AND WITNESS CONFIGURATIONS

These two additional records concern the surface of a blue sphere of radius 35 mm, not a tabletop contact disk. The sphere centers are $(0.550, 0, 0.460)$ m for bare-hand M03 and $(1.000, 0, 0.460)$ m for rod-assisted M06. Their horizontal center distances are 450 and 900 mm from the physical near edge at $x = 0.100$ m, rather than from the world origin, robot, or sphere's near surface. The sphere bottom is tangent to the physical tabletop at $z = 0.425$ m. Both retain the nominal 150-mm seated layout and external support.

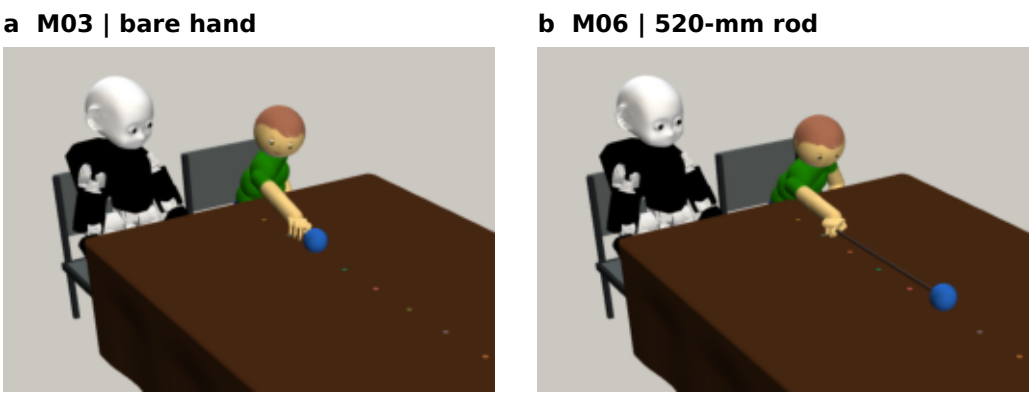


Figure 21: Archived blue-sphere static witnesses, distinct from the disk tasks in Figure 17. Both panels reuse the same fixed panorama camera and identical image crops/display scale. (a) Bare-hand M03 is the retained contact alternative, not the default ray-only posture. (b) M06 uses the earlier rod-assisted contact posture, not the benchmark's non-contact pointing state. Positive submillimetre gaps are taken from independent geometric records, not resolved from these pixels. Lean, turning, occlusion and original scene geometry remain visible. No new rendering, posture change or image retouching was performed.

The bare-hand M03 example is a contact configuration distinct from the benchmark's non-contact pointing pose. The rod-assisted M06 example likewise uses an earlier contact configuration rather than the benchmark pose. Each example is associated with its own saved geometry, camera, and validation results. These two configurations are used only to illustrate their respective physical protocols; neither replaces a benchmark stimulus.

S.2 TOLERANCES, POSTURE AND GRASP

For the bare-hand witness, the designated index-distal endpoint has gap $g = \|p - c\| - r_b$, where $r_b = 35$ mm. The recorded endpoint and full-capsule distances agree at 0.020000018 mm. The registered positive-gap tolerance is 0.05 mm; full-capsule penetration may not exceed 10^{-8} m. For the rod-assisted witness, the full physical tip sphere has radius $r_t = 5$ mm, so $g = \|c_{\text{tip}} - c\| - r_t - r_b = 0.055223415$ mm. This agrees with the native geometry query and passes the 0.10-mm gap tolerance; its numerical penetration allowance is 1 μm. Both observed gaps are positive. These are tolerance-based proximity witnesses, not measurements of force-bearing contact.

The bare-hand witness preserves the 22 right-finger coordinates of its registered bare-hand reference and the saved lower-body configuration. The tool-assisted witness preserves its accepted tool template's 22 right-finger coordinates, rod orientation in the palm and transverse placement, but permits translation along the rod axis. Its selected 4-mm station is the hand-origin reference, not the entire contact band. This differs from a search requiring the complete hand–rod transform to remain fixed; it does not demonstrate a causal benefit from regrasping. The physical rod remains 520 mm long. The seat values use the near edge minus the mean hip-flexion anchors: the bare-hand values

Table 25: Parameters of the two archived blue-sphere witnesses. Angles describe the actual state; total tilt is not a pure forward-lean angle. Dashes denote inapplicable quantities.

Quantity	Bare-hand M03	Tool M06
Surface gap / tolerance (mm)	0.0200 / 0.05	0.0552 / 0.10
MIMo hip-anchor seat distance (mm)	150.0017	150.0000
Observer hip-anchor seat distance (mm)	150.1798	150.0661
Actual total torso tilt (°)	30.4708	38.4595
Chest yaw (°)	47.0710	40.6640
Reported lateral tilt (°)	11.3123	3.6770
Hand-origin station from rod tail (mm)	—	4.000
Lateral grasp-contact band from tail (mm)	—	9.467–31.516
Non-target solid-table bound (mm)	3.200	3.295
Other geometry–sphere minimum (mm)	8.334	5.055

are reported directly; the tool-assisted values are reconstructed from archived body transforms and compiled joint-anchor constants without forward simulation.

Both protocols retain native joint/tendon constraints with 10^{-5}-rad acceptance tolerance, while removing the artificial 18° total-tilt ceiling and extra yaw, lateral-tilt and wrist gates. The archived audits pass the registered table, cloth, non-target sphere, support and abnormal-penetration checks. The table clearance floor is 3 mm; support penetration up to 1 mm and abnormal non-support penetration up to 0.2 mm are separate allowances. The tool-assisted protocol also checks geometric wrap and shaft clearance: maximum hand–rod penetration is 0.0239 mm and left-hand support penetration is 0.6525 mm. Neither posture is established as natural or robust. The bare-hand witness's chest-rotation joint has only 1.00005×10^{-7} rad (about 0.00000573°) remaining margin. The tool-assisted witness's minimum wrist margin is 2.419° and elbow flexion is 0.065°.

S.3 Scope relative to the original evidence

The old M03 certificate covers the registered disk task and Q_{18}, including its additional pelvis assumptions. It cannot be applied to either blue-sphere task. For example, the M03 sphere reaches world $x = 515$ mm, whereas the old disk starts at 542.5 mm; the old $x < 540$ mm enclosure is not a sphere exclusion. The new postures also exceed 18°. The expanded-domain disk M03/M06 unknowns remain intact. Changing target surface/height, posture and permitted grasp adjustment together is not a controlled comparison of object size. Each new witness establishes one admissible configuration within its tolerances, not a farthest reach, a continuous workspace or an exclusion of subsequent targets.

The archived eye/ray checks concern these saved states only. They retain partial silhouette occlusion and do not establish VLM recognition. Blue material input remains `#2463BF`, but these scene records use specular/shininess values 0.02/0.04, unlike the zero-valued standalone overview settings in Appendix G; the twelve-appearance figure is unchanged. The tool-assisted scene already matches its visible cylinder to the physical shaft and exposes the existing tip sphere. This is a recorded scene change, not retouching in this paper, and does not remove the old disk scene's disclosed rod-rim overlap. Its retained visual cloth top overlaps the sphere bottom by about 0.025 mm; the physical sphere–table tangent and hanging-cloth checks are distinct.

These examples are supported by saved-state geometry and validation records, not by new collision checks or dynamic execution in the present evaluation. Force-bearing contact, sustained contact, grasp stability, and disturbance robustness remain unevaluated. The examples do not establish validity for other shapes or colors, or recognition of their images. Static checks and VLM results for the separate benchmark collection are described in Appendix H and Section 5.1.

The benchmark M03 bare-hand and M06 held-tool templates in both layouts have their own non-contact pointing configurations. Their surface gaps and interaction labels are specific to those configurations and cannot be copied from the historical contact examples. Neither the historical witnesses nor formal ray selection establishes the maximum reachable marker.

Table 26: Historical evidence and formal saved states have different identities and scopes. Formal ray status is not unreachability.

Source / target	Evidence in its own protocol	Relation to benchmark stimuli
Historical disks	M02 dynamic; M05 static; M03 conditional exclusion in Q_{18}; expanded M03/M06 unknown	Original disk geometry only; no automatic transfer to new objects
Bare-hand M03 sphere	Contact configuration; 0.0200 mm gap	Separate contact alternative; formal bare-hand M03 is ray in both layouts
Tool-assisted M06 sphere	Contact configuration; 0.0552 mm gap	Separate contact alternative; formal held-tool M06 is ray in both layouts
Formal collection	576 static scenes; 144 templates	Separate poses and geometry checks; no dynamic validation or reachability certification