MENTISOCULI-3D: A VERIFIED BENCHMARK FOR 3D STATE TRACKING UNDER PRESCRIBED ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating spatial prediction requires both a correct terminal label and evidence that the supplied observations determine it. We introduce MENTISOCULI-3D, a procedural benchmark for tracking 3D state under prescribed actions. Five task families cover articulated markers, hidden token identities, support release, counterfactual branches, and combined root and carousel rotations. Each sampled root yields two questions with identical text and different verified answers. The generator checks initial-view evidence, marker visibility, independent kinematic execution, and numerical stability of physical outcomes. Matched inputs vary access to intermediate images, exact states, and terminal coordinates. The artifact also contains 4,800 distribution-shift and long-horizon questions, diagnostic banks, and an interactive human-review interface that records visual assistance. We evaluate seven open checkpoints and GPT-6 Astra on the 4,000-question core. Open-model accuracy ranges from 22.35% to 33.40% under a 2,048-token response cap. In a separate Codex configuration with high reasoning effort, Astra scores 98.63 [98.25, 98.98]%, placing this configuration near the benchmark ceiling. Two open-model observation panels achieve high terminal-coordinate readout but show no clear average benefit of feedback over repeated initial images. These results support reporting information controls and response validity alongside accuracy. They characterize behavior on the generated tasks, without establishing an internal simulation mechanism. Two human reviewers checked approximately 10% of the benchmark dataset and certified the correctness of the reviewed subset.

Keywords: multimodal evaluation; 3D state tracking; procedural benchmarks; counterfactual reasoning; evaluation validity.

1 INTRODUCTION

A red marker is attached to the end of a two-link mechanism. The root turns about its current local x axis, the child hinge bends, the root turns about its current local z axis, and then about a fixed world axis. Which floor quadrant lies below the marker? The initial image and the action program jointly determine the answer. Naming the marker's starting quadrant, ignoring the distinction between local and world axes, or applying the actions in a different order can each produce a plausible but incorrect response. Figure 1 illustrates all five families; Appendix J shows construction and action replays. A related question about a covered token is meaningful only if an earlier view revealed which token occupied the compartment.

MentisOculi motivates controlled evaluation of repeated visual manipulations (Zeller et al., 2026). Extending this approach to three dimensions adds moving reference frames, articulated joints, and contact, but also creates an evaluation problem: simulator access to a hidden state does not imply that the model's observations determine the answer. Low accuracy is difficult to interpret if the initial geometry is ambiguous, the queried marker is too small, or the physical outcome changes under a minor numerical perturbation.

MENTISOCULI-3D treats the related questions and their verification records as the unit of construction. Two questions share their complete textual prompt and option order, while a change in the initial state changes the correct semantic answer (Figure 10). A deterministic text-only system must give the same answer to both and therefore cannot answer the pair entirely correctly. The generator

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.

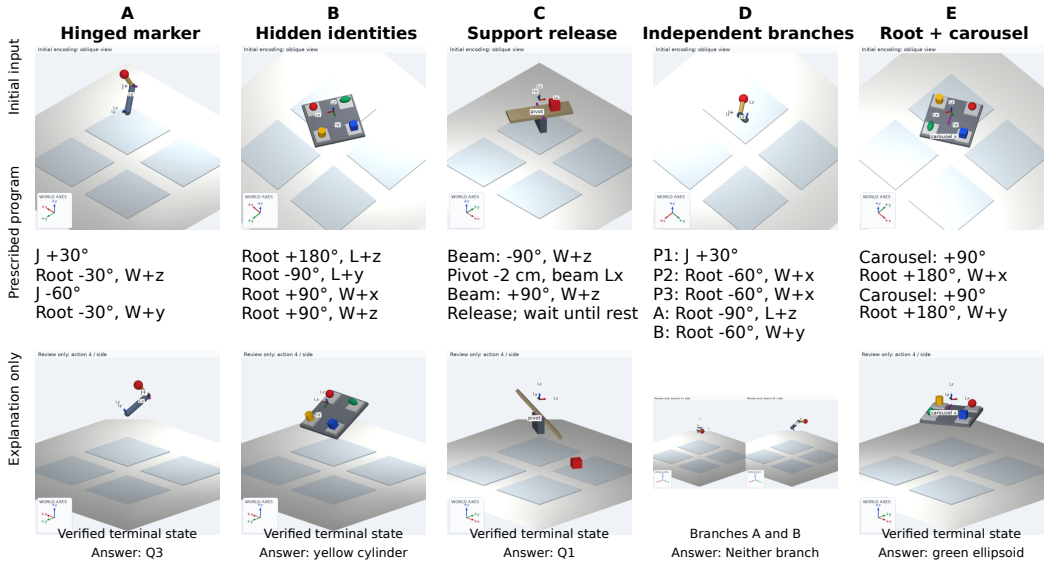


Figure 1: **Five scenarios require tracking different parts of a 3D state.** Initial views and prescribed actions determine marker locations, token identities, and release outcomes. The bottom row reveals verified endpoints for explanation; these views are withheld from open-loop (OL) inputs. W and L denote fixed-world and current-local axes.

balances semantic answers and display letters jointly, checks the evidence supplied to the model, and preserves the paired structure in splits and analysis. Prescribed actions remove planning and policy selection from the primary task.

The dataset contribution is a construction protocol that links each question to its observed evidence, answer-changing twin, execution trace, and acceptance records. The evaluation contribution is a matched set of information conditions, pair scores, and response-validity controls. Together they support comparisons that a single aggregate spatial-reasoning score cannot provide. We implement these components for three mechanism grammars and evaluate seven open checkpoints and GPT-6 Astra on the full core. An interactive review tool lets people apply each action and inspect its result; responses using this assistance are separated from unaided performance. All reported model results concern the generated distribution and recorded inference settings.

2 Related Work

Visual simulation and spatial updating. MentisOculi studies procedural visual manipulation, including hinge folding, and tests intermediate visualizations (Zeller et al., 2026). STARE covers 2D/3D transformations, cube-net folding, and visual simulation aids (Li et al., 2026). These works already establish multistep transformations and visual assistance as evaluation tools. Our contribution couples every 3D scene to an answer-changing twin with identical text, verifies both variants, and preserves that dependency in scoring. Repeated-initial-image controls distinguish updated state information from simply adding image slots. Exact initial, checkpoint, and terminal states then probe different information requirements on the same questions.

MindCube evaluates limited-view spatial modeling, while Cube Bench studies sequential cube reasoning (Wang et al., 2026; Anand & Shareghi, 2025). SpatialAct, Theory of Space, and SpatialWorld examine action-based reasoning or active exploration (Liu et al., 2026; Zhang et al., 2026a; Gao et al., 2026). We supply a fixed program: the solver predicts its outcome without selecting an exploration or control policy. This restriction makes paired state changes and matched information conditions possible, at the cost of narrower task coverage.

Physical prediction and input-grounded validity. CLEVRER, CoPhy, and CATER establish controlled temporal and counterfactual prediction (Yi et al., 2020; Baradel et al., 2020; Girdhar

& Ramanan, 2020); Physion++ and IntPhys 2 evaluate physical understanding (Tung et al., 2023; Bordes et al., 2025). Our physical family discloses mechanical parameters and verifies a restricted support-release outcome under integration and image-position perturbations. MindEdit-Bench studies hypothetical object edits in photographs (Yu et al., 2026). ReVSI checks spatial answerability against the actual supplied views and provides visibility metadata (Zhang et al., 2026b). We apply this input-grounded principle during procedural construction: finite pose checks, visible-target masks, boundary margins, and paired closed-image checks connect each label to its available evidence. These checks remain weaker than human validation or global identifiability. The distinctive contribution is the resulting verified, paired evaluation protocol across three mechanism grammars (Table 51), rather than 3D rendering or intermediate visualizations alone.

3 Benchmark construction

3.1 States, actions, questions, and paired roots

For a sampled mechanism θ, let S_0 be the initial state, $a_{1:T}$ a prescribed program, F_θ its transition function, and G_q an executable query. A question has label

$$S_t = F_\theta(S_{t-1}, a_t), \qquad y = G_q(S_T) \in \{0, 1, 2, 3\}. \tag{1}$$

Kinematic states contain a root rotation and joint angle; the physical grammar also uses pivot position and complete MuJoCo integration state. The supplied images show the initial state from an oblique and an overhead camera. In the hidden-token tasks (B and E), these two views reveal the tokens and their positions. A third image shows the carrier with its opaque lids closed, hiding the tokens without changing their positions. The action horizon counts only the prescribed puzzle actions; it excludes the initial viewing stages, closing the lids, and the simulator's internal time steps. For Task D (counterfactual branches) in the core test set, the action horizon equals the number of shared actions plus one alternative final action. Each branch starts from the same state S_{T-1} reached after the shared actions. The answer function G_q evaluates the two final states obtained by applying each alternative last action separately to S_{T-1}.

A root contains two variants with different semantic answers and identical language, including the action program and option order. A and D vary the root orientation while keeping the initial hinge angle fixed. B and E exchange the two identities needed to change the answer. C reflects the block's local offset. Both questions in the matched pair must individually pass all dataset acceptance checks described below; accepting one does not automatically qualify the other. We group two consecutive matched pairs into a balancing group of four questions, so that each of the four possible answer meanings is correct exactly once. For example, a quadrant task has one correct answer each of Q1, Q2, Q3, and Q4. All four questions share one randomly shuffled option order, so each answer letter (A–D) is also correct once. The option order varies across groups. The generator uses separate sequences of random numbers to choose the scene geometry, visual appearance, action sequence, question, and answer-option order. Both questions in each matched pair belong to the same dataset partition (pilot, development, or test); a pair is never divided across partitions.

Computing the answer. We compute labels from executed geometry. For a token with carrier-local center u_i, its final world center is $p_i = o + R_T R_z(\phi_T) u_i$, where o is the hub's world position. The rotation matrix R_T maps root-local coordinates to world coordinates after T actions; $R_z(\phi_T)$ rotates the carousel by its final joint angle ϕ_T about the root's local z axis. Apply the carousel rotation first, then the root rotation, then add the hub position. For the rigid carrier (Task B), $\phi_T = 0$ and $R_z(0)$ is the identity matrix. Quadrant 1 (Q1) of the fixed world x–y plane contains the unique token whose final center satisfies $p_{i,x} > 0$ and $p_{i,y} > 0$. The other quadrants use the corresponding sign pairs. We then map the correct token to its displayed answer letter. For the marker (Task A) and block (Task C), the answer is the quadrant containing its center after all actions, using the fixed world x and y axes. Task D checks whether the marker ends on the positive world-x side ($x > 0$) in branch A only, branch B only, both branches, or neither. Rendering produces the input images; executing the actions and evaluating the question produces the correct answer. Appendix A gives the task-specific equations and the ten-step generation procedure.

Bank	Roots	Questions	Purpose / action horizon
Pilot	200	400	Development checks; $1, 2, 4, 8$
Development	520	1,040	Construction and human review; $1, 2, 4, 8$
Core test	2,000	4,000	Primary model comparison; $1, 2, 4, 8$
Four shift banks, combined	2,000	4,000	Geometry, camera, appearance, magnitude; $4, 8$
Long-horizon stress	400	800	Longer programs; $12, 16$

Table 1: Main dataset inventory. Roots have two answer-changing variants. Matched observation conditions reuse questions rather than enlarging these counts. The 400-question human-review allocation is a development subset. Auxiliary diagnostic banks are reported separately.

3.2 Acceptance checks

Keeping object shape and dimensions unchanged, we compare the image positions of identified scene points (landmarks) across the allowed initial poses. We reject a scene if a different root orientation or joint angle places every checked landmark within two pixels of its original position in both views. This check distinguishes only the generator's allowed poses, not all possible 3D arrangements. In Tasks B and E, the initial views must reveal all four tokens: the red sphere, blue cube, yellow cylinder, and green ellipsoid. Within each matched pair, the images with the lids closed must contain identical image data.

Each marker or token requires a qualifying view with at least 400 visible pixels, a 12-pixel bounding-box extent, and 60% of its isolated silhouette; masks exclude text overlays. Final centers must stay at least 2 cm from the answer boundaries: $x = 0$ and $y = 0$ for quadrant questions, or $x = 0$ for Task D. These checks limit occlusion and boundary ambiguity but do not establish human readability.

We compute rotations with Rodrigues' rotation formula and check them using quaternions and a separate MuJoCo model of connected parts and joints, including both branches of Task D. For C, MuJoCo 3.13.0 (Todorov et al., 2012) uses 0.5 ms time steps. The answer must remain unchanged with 0.25 and 0.125 ms steps, 16 reproducible trials varying block pose and contact parameters, and four combinations of block/pivot position errors equivalent to ± 2 image pixels. Appendix A.4 explains these robustness tests in detail. Runs must settle within five seconds with less than 2 mm penetration. Restored checkpoints must match uninterrupted simulation. Appendix A specifies tolerances, perturbations, torque margins, and regeneration.

3.3 Splits and auxiliary banks

Table 1 separates the main banks from reused observation conditions. The core has 100 roots per task/horizon cell; development has 26, preserving four-way balance. All paired variants stay together. Test generation and acceptance are fixed before inference, and test seeds and labels are excluded from model requests. No question is removed because of a model's answer.

The four extension banks vary geometry, oblique camera, neutral appearance, or primitive action magnitudes independently. The camera shift retains the overhead anchor; the program shift does not hold out an entire composition grammar. Stress extends the action horizon to 12 and 16. Separate evidence-sufficiency, branch-consistency, and hold-padding diagnostics retain their own denominators and controls. Padding preserves endpoints; the active-program control changes the computation. Single-action hold sensitivity does not certify minimum reasoning depth. Detailed ranges and inventories appear in Appendix A.

4 Evaluation protocol

4.1 Matched information conditions

Open-loop (OL) supplies initial images and all actions in a fresh model context. Visual feedback (FB) adds intermediate images; closed lids still hide tokens. Repeated-view control (REP) repeats initial images in matching slots. Exact initial state (S0) replaces images with numeric state information. Midpoint reset (RESET) adds the exact halfway state, retaining all actions. Final-state

readout (FINAL) supplies final coordinates without images or actions, testing interpretation rather than action tracking. At $T = 1$, identical OL/FB/REP/RESET requests reuse one response. Task D feedback ends before branching. All conditions use the same questions and option order. Differences in information, precision, and representation prevent attribution to a single internal failure cause. Table 53 specifies all inputs.

4.2 Models, inference, and scoring

The open-model panel comprises Qwen3-VL-4B/8B-Instruct, Qwen3.5-4B/9B, InternVL3.5-8B, and SmolVLM2-2.2B-Instruct (Bai et al., 2025; Qwen Team, 2026a; Wang et al., 2025; Hugging Face Smol Models Research, 2025). Qwen3.8-27B-FP8 (Qwen Team, 2026b) completes this seven-model panel. All seven cover the 4,000-question core. Their original matched 400-question comparison retains five complete balancing groups per task/horizon cell. Open-model inference uses vLLM 0.29.0 (Kwon et al., 2023) on one RTX 5090 with 32 GB nominal memory. Revisions and serving settings are recorded per run. SmolVLM2 uses a SmolLM2 decoder; InternVL uses a Qwen3 backbone, so this panel is dominated by Qwen.

GPT-6 Astra also answers all 4,000 OL questions through Codex CLI 0.156.1, with `model_reasoning_effort="high"` and a fresh ephemeral session per question. Only public puzzle inputs are supplied; tools are excluded. Its output budget, sampling, and provider snapshot differ from the open-model contract (Appendix K); scoring uses the same simulator labels.

The shared instruction requests at most 150 words of reasoning and a terminal `FINAL: X` answer. Open-model generation uses temperature zero, a 2,048-token output cap, and disabled native thinking where the checkpoint template supports that switch. The word limit is a prompt instruction, not a tokenizer constraint; output lengths and truncations are reported. A syntactic parser accepts one case-insensitive terminal answer tag with an optional period. Multiple tags, trailing explanations, and missing tags are invalid. Invalid or truncated responses are scored as produced, without semantic repair. Transport/server errors instead make a run incomplete and are repaired before reporting a complete score. Every open checkpoint also receives OL and FINAL on the same 400-question subset with an instruction to return only the answer tag and a 32-token cap. Instruction and budget change, so the comparison does not isolate reasoning length. The 400 questions form a nested subset of the observation bank. For its two models, a post-hoc FINAL comparison uses these same questions under both response contracts, without additional inference. The 27B follow-ups vary the output cap and the model-card sampling bundle on 100 matched eight-action questions. Their allocation rules and settings are recorded in Appendix C. The additional two-model experiments add separate fixed-instruction cap and repeated-sampling comparisons.

Let $c_{r,v}$ indicate correctness for variant v of root r. The primary score macro-averages the 20 task/horizon cells. Pair accuracy replaces each root's mean correctness with $c_{r,0}c_{r,1}$. Uniform independent guesses have expected question accuracy 25% and pair accuracy 6.25%; a deterministic system that uses only the shared prompt has pair accuracy zero. We compute percentile 95% intervals from 10,000 stratified bootstrap draws, resampling the two-root balancing groups so both counterfactual and quota dependencies remain intact. Condition differences use paired draws over the same groups. Auxiliary scores are reported separately with their own denominators and trivial baselines. Intervals are pointwise and are not adjusted for the number of comparisons.

5 Experiments and results

5.1 Verification, model coverage, and response validity

All 4,000 core labels pass independent post-generation execution checks, including both counterfactual suffixes and the physical checkpoint replays. Acceptance records retain the physical perturbation runs and visibility checks. A privileged executor reproduces the labels, while development-fitted language probes and image/program removal controls test simpler alternatives. Full numerical audits and shortcut comparisons appear in Appendix H.

Across the seven open checkpoints, full-core accuracy ranges from 22.35% to 33.40%. Qwen3.8-27B-FP8 scores 31.63%, with 57.05% invalid responses. Astra with high reasoning scores 98.63% [98.25, 98.98], with 97.35% pair accuracy and no invalid answers (Figure 2; Table 54). This result

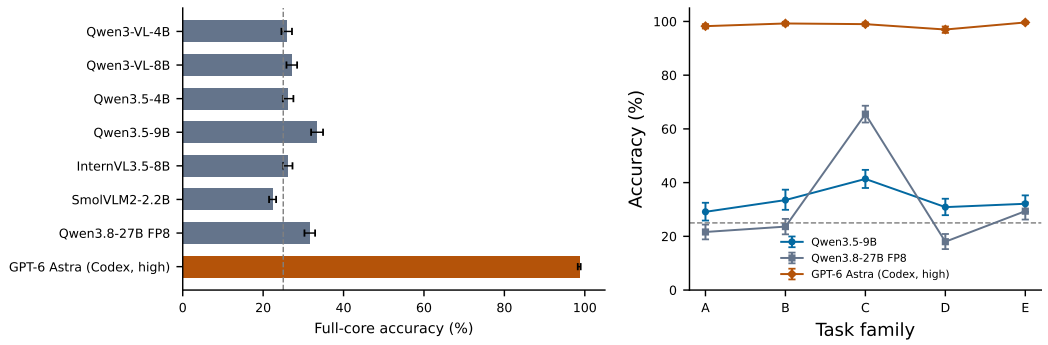


Figure 2: **High-reasoning Astra nearly solves the core; open-model scores remain much lower.** Left: accuracy on the same 4,000 questions. Right: task profiles for Astra and two open models. Astra uses a separate Codex high-reasoning contract; the open models use the capped brief contract. Bars show pointwise 95% balancing-group intervals; dashed lines mark 25% chance accuracy. Invalid outputs count as wrong.

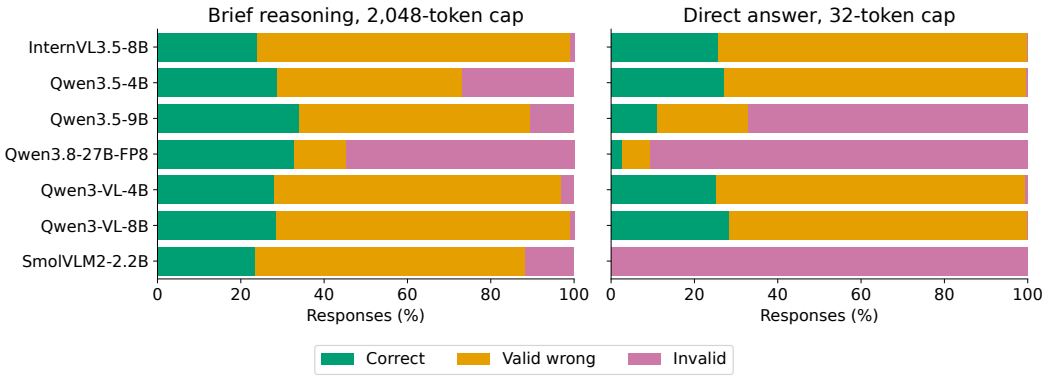


Figure 3: **Invalid output explains part of the measured failure.** Correct, valid-wrong, and invalid responses are shown separately on the same 400 questions. The direct contract changes both the instruction and token cap; its effect cannot be attributed solely to reasoning length.

places Astra near the ceiling of these task families under its recorded settings. The configurations differ in reasoning budget, image processing, and harness, so this is not a compute-matched estimate of a model-identity effect.

Response validity changes the interpretation of these scores (Figure 3). Under the direct 32-token contract, SmolVLM2 emits a bare option on all 400 OL questions, yielding zero credit under the frozen tag parser. Qwen3.5-9B loses 23.00 percentage points relative to its brief contract (95% interval $[-28.50, -17.50]$). These interventions change both instructions and output caps. Separate fixed-instruction cap experiments are reported below; none of these scores estimates the best achievable performance of a checkpoint.

5.2 What information helps?

Exact terminal coordinates yield 95.38% and 98.75% accuracy for Qwen3.5-9B and InternVL3.5-8B on their matched 800-question panels. Supplying exact initial state or a halfway checkpoint produces much smaller scores (Figure 4). Terminal readout therefore checks a capability that can succeed while full state propagation remains difficult. The conditions change representation, precision, and remaining computation; their differences cannot be added into independent perception and reasoning error rates.

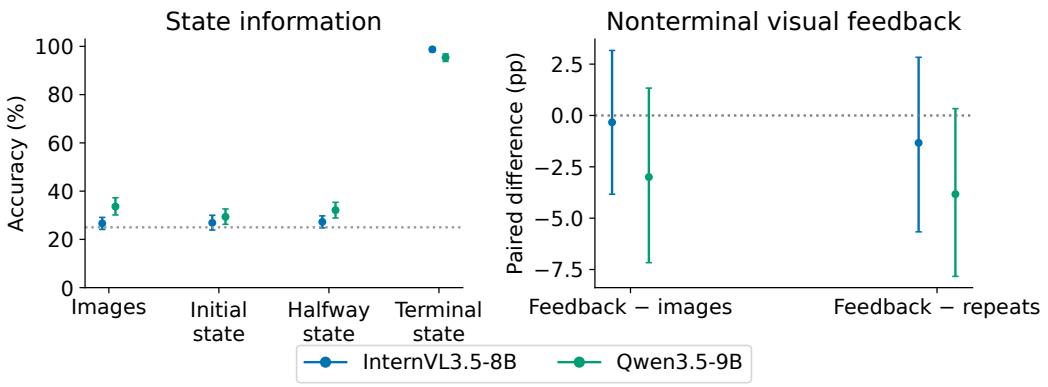


Figure 4: **Terminal readout succeeds while intermediate feedback gives no reliably positive aggregate gain.** Left: the same 800 questions under each state-access condition. Right: paired feedback contrasts on $T > 1$, excluding identical one-step requests. Error bars are pointwise 95% intervals.

Nonterminal feedback shows no reliably positive aggregate effect over either OL or repeated initial images (Figure 4). The task-specific plots in the appendix show why the repetition control matters: a gain relative to OL need not establish a gain from updated state information. We exclude the identical one-action requests from this comparison. Covered token identities remain covered in FB, so the result concerns this particular feedback protocol.

5.3 State trajectories, frame equivalence, and reasoning aids

We extend the two-model analysis with a prospectively frozen exploratory protocol on the same matched panel. Coordinate and tool conditions disclose static mechanism dimensions and hub location; a geometry-only final-answer control receives the same calibration. Initial orientation, joint angle, and token assignment remain to be inferred. Independent prefix queries and rollouts that receive the immediately preceding model prediction are evaluated separately on 160 questions. Exact checkpoints are assigned before observing errors.

The frame diagnostic retains 44 questions in 11 complete source balancing groups. Each two-action program has a verified local-axis equivalent and an order-swapped program with a different answer. Equal-length three-action controls swap commuting parts of a split rotation. Numeric local-axis coefficients can disclose directional information. Each accepted group retains four-way label balance, but task coverage is small and uneven; this limits the frame comparison.

On all 400 questions, we test state-table prompting, two fixed development examples, output caps, four stochastic samples, and a fixed executor operating on model-estimated initial state. A separate terminal-state query compares the readout of predicted coordinates with the model's final answer. We retain all raw responses, request descriptions, tool results, and sampled outputs. Appendix I specifies the prompts, tolerances, information boundaries, and complete results. The original terminal-state contract hits the output cap on 399 and 0 of 400 questions for Qwen3.5-9B and InternVL3.5-8B, respectively. Such incomplete responses cannot isolate geometric error. We therefore report a separate post-hoc diagnostic that requests immediate JSON and constrains its schema; it changes both instruction and decoding. Complete terminal JSON reaches 100.00 [100.00, 100.00] and 100.00 [100.00, 100.00]%, but schema completeness alone does not establish a valid pose or correct coordinates.

Before any action, guided independent coordinate success is 10.00 [0.00, 20.00] and 0.00 [0.00, 0.00]%. At the eight-action endpoint it is 0.00 [0.00, 0.00] and 0.00 [0.00, 0.00]%; supplying the preceding prediction yields 0.00 [0.00, 0.00] and 0.00 [0.00, 0.00]% (Figure 5). The appendix reports unconditional validity, continuous error among valid states, and coordinate–answer joint outcomes for both response contracts. Initial-state errors limit what endpoint failure can establish about error accumulation.

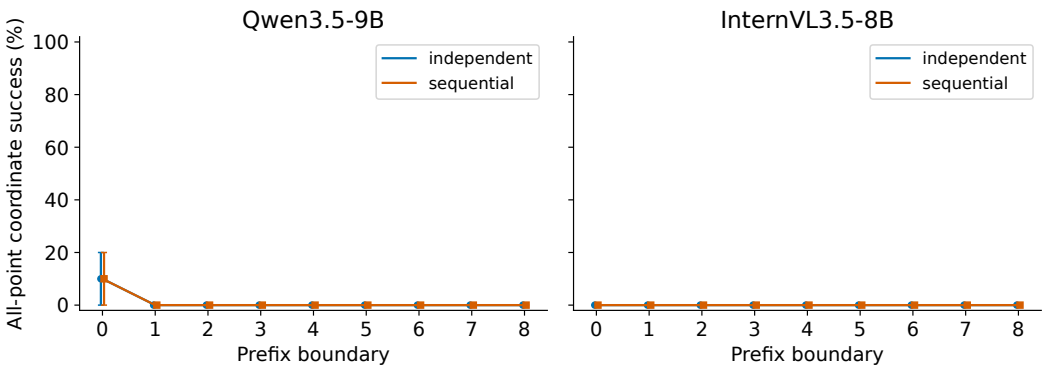


Figure 5: **Accurate initial coordinates remain rare under guided JSON.** The post-hoc JSON contract fixes the schema without supplying true coordinates. All named points must be within 2 cm after geometric validity checks. Both protocols use the same $T = 8$ cohort (40 questions); sequential queries receive the preceding prediction and full prefix. Static dimensions are disclosed, D endpoints are averaged within item, and intervals are pointwise 95%.

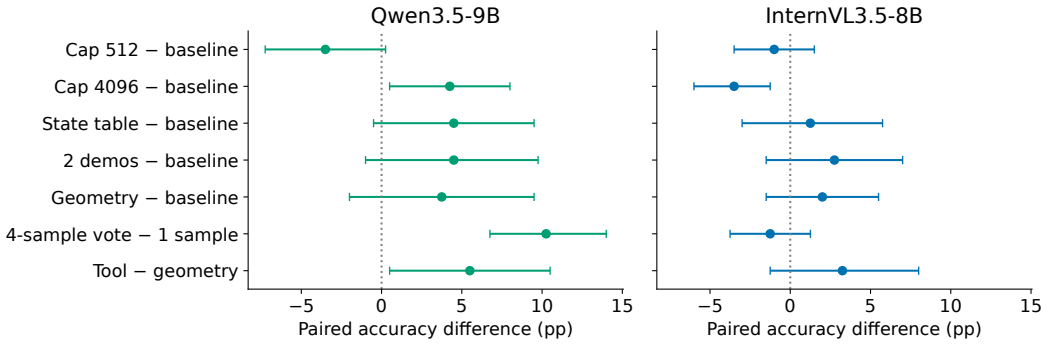


Figure 6: **Reasoning aids have different effects under matched controls.** All original-contract contrasts use the same 400 questions. Cap and prompting changes use a fresh brief baseline; voting uses its first stochastic sample, and tools use geometry-only calibration. Voting takes four calls and tools two. The first-valid voting control and executor failure rates accompany these contrasts in the text and appendix. Intervals are pointwise 95%.

At the assigned last-nonterminal checkpoint (80 questions), exact state improves accuracy over matched scaffolding by 12.50 [1.25, 23.75] and 8.75 [1.25, 16.25] percentage points. Midpoint intervals include zero; the two checkpoint categories have different memberships, so their contrast does not isolate checkpoint timing.

Figure 6 compares reasoning aids with matched controls. Four-sample voting changes accuracy relative to its first stochastic sample by 10.25 [6.75, 14.00] and -1.25 [-3.75, 1.26] percentage points. Relative to the first valid answer among those same four samples, the change is 3.25 [0.75, 6.00] and -1.25 [-3.75, 1.26] points. This retrospective control uses all four recorded calls; it does not measure an early-stopping policy.

The original numeric-tool condition fails to execute on 89.25 [86.50, 91.75] and 40.00 [36.50, 43.75]% of questions. Under guided JSON, failure rates are 19.75 [17.00, 22.75] and 41.75 [38.00, 45.50]%, and tool-minus-guided-answer-only accuracy is 0.25 [-3.25, 3.75] and 1.50 [-0.75, 3.75] points. Tool errors are passed to the second model call; a final-answer gain alone therefore does not demonstrate correct execution.

5.4 TRANSFER AND DIAGNOSTIC CONTROLS

The geometry, camera, appearance, and action-magnitude banks produce accuracies near the uniform reference for the two evaluated models; the longer stress bank is also difficult. These banks contain independent scenes, so their differences from the core reference are not causal effects of rerendering a fixed episode. Paired-scene, branch, evidence-sufficiency, and padding plots provide additional controls. Agreement after reconvergence is reported with correctness because a constant prediction can agree without tracking state. Hold-padding preserves endpoints, whereas the active-program control changes the required computation. Appendix H reports every condition and its uncertainty.

6 HUMAN REVIEW AND DATASET MAINTENANCE

Two human reviewers checked approximately 10% of the benchmark dataset and certified the correctness of the reviewed subset. This inspection complements the automated checks. We have not measured unaided human accuracy, completion time, or structured readability ratings.

Reviewers submit answers, confidence, timing, and visibility/clarity/solvability feedback. They can apply actions and inspect three explanatory views. Logged assistance and prior replay exposure exclude responses from unaided accuracy and clean timing summaries. Voluntary replay does not establish causal benefits. An append-only queue records ambiguity reports and evidence for group-level revisions or exclusions in future versions, preserving the tested bank. Appendices D and E document review procedures and the dataset.

7 LIMITATIONS AND RELEASE STATUS

The distribution has three compact mechanism grammars, one child joint at most, fixed rules, and templated language. Joint increments separate algebraically from root updates, and C preparation reduces to net yaw and pivot changes. Every accepted C quadrant label also agrees with a privileged torque-sign rule, so its coarse outcome need not be obtained by full trajectory simulation. The tasks do not cover general contact planning, arbitrary objects, or real photographs. Physical outcomes depend on one engine and contact configuration. Finite projection checks hold geometry fixed and cannot establish global inverse-graphics identifiability. Nine core questions have partially occluded local-x labels in both initial views; we retain and flag them for readability review, and none belongs to the matched 400.

The open-model panel is mostly Qwen. All eight models cover the core, but only two open models cover all interventions and shifts. Astra's near-ceiling score limits discrimination among stronger systems. Its high-reasoning run lacks a matched low-effort control and a frozen checkpoint identifier. Serving settings do not establish each model's best achievable score. Coordinate queries supply metric geometry and change the output format; frame rewrites add numeric directions. These interventions and the small frame bank do not establish general internal simulation. Intervals are pointwise and conditional on the recorded samples and serving runs.

8 CONCLUSION

MENTISOCULI-3D evaluates 3D state tracking with verified paired questions. Astra nearly solves the core at high reasoning effort; seven capped open-model runs score much lower. Two observation panels show strong terminal readout but weak initial-view prediction, with no clear average feedback benefit. These findings support reporting pair accuracy, invalid outputs, and information controls alongside accuracy.

REPRODUCIBILITY STATEMENT

The artifact records generator/model revisions, requests, responses, certificates, and analysis settings. Private labels and raw responses remain local; disclosed cluster scores reproduce intervals. Appendix C describes commands and information boundaries. Quantitative text, tables, and figures are generated from recorded summaries.

ETHICS STATEMENT

The scenes use synthetic primitive geometry and contain no photographed people or personal data. Two humans reviewed approximately 10% of the benchmark dataset for correctness. Recruitment for the separate participant-performance study requires the applicable consent and institutional review determination. Response files use assigned participant numbers and remain private. The coarse physical tasks are unsuitable for safety-critical simulation, and assisted review accuracy must not be reported as unaided human performance. Licensing and data access are documented in Appendix E.

STATEMENT OF GENERATIVE AI USE

An AI coding assistant supported literature review, benchmark implementation, tests, experiment orchestration, analysis code, and manuscript drafting. We disclose this assistance separately from the evaluated model responses; the numerical results come from the retained inference records and executable scoring code.

REFERENCES

Dhruv Anand and Ehsan Shareghi. Cube bench: A benchmark for spatial visual reasoning in mllms, 2025. URL https://arxiv.org/abs/2512.20595. arXiv: 2512.20595.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631. arXiv: 2511.21631.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics, 2020. URL https://arxiv.org/abs/1909.12000. arXiv: 1909.12000.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849. arXiv: 2506.09849.

Hongcheng Gao, Hailong Qu, Jingyi Tang, Jiahao Wang, Zihao Huang, Hengkang Qiao, Shihong Huang, Junming Yang, Yi Li, Hongyixuan Yuan, Wenjie Li, Bohan Zeng, Wenbo Li, Bo Wang, Jianhui Liu, Olive Huang, Haoyang Huang, Wentao Zhang, Guoqing Huang, Nan Duan, and Yinpeng Dong. Spatialworld: Benchmarking interactive spatial reasoning of multimodal agents in real-world tasks, 2026. URL https://arxiv.org/abs/2606.09669. arXiv: 2606.09669.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning, 2020. URL https://arxiv.org/abs/1910.04744. arXiv: 1910.04744.

Hugging Face Smol Models Research. SmolVLM2-2.2B-Instruct model card, 2025. URL https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct. Official model card; revision 482adb537c021c86670beed01cd58990d01e72e4; accessed 10 September 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fbGmSV6tUw.

Tianhui Liu, Jie Feng, Zhiheng Zheng, Shengyuan Wang, Yiming Guo, Yanxin Xi, Hangyu Fan, Yong Li, and Pan Hui. Spatialact: Probing spatial reasoning-to-action capabilities of vlm agents in 3d scenes, 2026. URL https://arxiv.org/abs/2605.31148. arXiv: 2605.31148.

Qwen Team. Qwen3.5-4B model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-4B. Official model card; accessed 10 September 2026.

Qwen Team. Qwen3.8-27B-FP8 model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.8-27B-FP8. Official FP8 checkpoint; revision 017b9c7af6b5689d5dd426a76e0bc077eb5ca20a; accessed 10 September 2026.

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.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel M. Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel L. K. Yamins, Judith Fan, and Kevin A. Smith. Physion++: evaluating physical scene understanding that requires online inference of different physical properties. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

vLLM contributors. Reproducibility. vLLM documentation, 2026. URL https://docs.vllm.ai/en/latest/usage/reproducibility/. Accessed 10 September 2026.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026. URL https://arxiv.org/abs/2506.21458. arXiv: 2506.21458.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265. arXiv: 2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171. arXiv: 2203.11171.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. Mindedit-bench: Benchmarking object-level counterfactual spatial reasoning in vlms from in-the-wild photos, 2026. URL https://arxiv.org/abs/2607.00491. arXiv: 2607.00491.

Jana Ricarda Zeller, Thaddäus Wiedemer, Fanfei Li, Thomas Klein, Prasanna Mayilvahanan, Matthias Bethge, Felix A. Wichmann, Ryan Cotterell, and Wieland Brendel. Mentisoculi: Revealing the limits of reasoning with mental imagery. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sxvuK2x3eA.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration?, 2026a. URL https://arxiv.org/abs/2602.07055. arXiv: 2602.07055.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X Chang. ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3d reasoning. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=FTC7najzZT.

A TASK DEFINITIONS AND GENERATION RECIPE

A.1 COORDINATE CONVENTIONS AND STATE TRANSITIONS

The world is right-handed, with $+z$ upward and a fixed hub at $(0, 0, 0.5)$ m. Root axes are labeled Lx, Ly, and Lz. An inset depicts the fixed world axes Wx, Wy, and Wz without allowing foreground geometry to occlude them. The four floor quadrants are Q1=$(+x, +y)$, Q2=$(-x, +y)$, Q3=$(-x, -y)$, and Q4=$(+x, -y)$. Membership is determined by the object's center projected vertically onto the floor; height is ignored. Positive angles obey the right-hand rule about the named positive axis: a positive rotation is anticlockwise when viewed from the positive-axis tip toward the root origin, and clockwise from the opposite direction. For 180 degrees both signs have the same endpoint. Local axes use their current orientation. The camera never transforms the object.

Let R map root-local vectors to world coordinates and let $R_k(\alpha)$ denote a primitive rotation about coordinate axis k. A world-axis action updates $R \leftarrow R_k(\alpha)R$; a root-local action updates $R \leftarrow RR_k(\alpha)$. A child-joint action updates its relative scalar angle. The corresponding world point is computed from the resulting root and joint transforms. These conventions are repeated in every rule card and tested with noncommuting rotations. No Euler-angle addition is used to execute a program.

Algebraic simplifications. Let $L_1, \ldots, L_p$ be the local root rotations in chronological order and $W_1, \ldots, W_q$ the world root rotations in chronological order. Then

$$R_T = (W_q \cdots W_1)R_0(L_1 \cdots L_p), \qquad \theta_T = \theta_0 + \sum_j \Delta\theta_j.$$

Left and right multiplication commute with one another; order can still matter within either ordered product. Relative joint increments commute with root updates in these single-joint grammars. E therefore combines hidden identity and two types of update with algebraically separable controls. For C, clamped preparation reduces to total yaw and net pivot translation, followed by one passive release simulation. Accepted intermediate joint/pivot limits remain part of the construction rules. These simplifications reinforce that action count and single-action hold sensitivity measure properties of the prescribed program, rather than its minimum required computational depth.

For the accepted C bank, a privileged analytical rule predicts the quadrant of

$$\operatorname{sign}(m_c x_c - (m_b + m_c)p)\, R_z(\psi_T)e_x,$$

where x_c is the initial block offset, p the prepared pivot offset, and ψ_T the prepared yaw. The masses and prepared values are obtained from exact initial geometry and the action sums. Agreement with the stored nominal labels is measured after generation; it is never used to accept a question. This control characterizes the coarse outcome query, while the simulator still supplies the trajectories and numerical stability checks.

Initial articulated poses use $R = R_z(\psi)R_y(\beta)R_x(\gamma)$, where $\psi \in \{30, 60, 120, 150, 210, 240, 300, 330\}$ degrees, $\beta \in \{-45, 0, 45\}$, and $\gamma \in \{-30, 0, 30\}$. The initial child angle is in $\{-60, 0, 60\}$, giving 216 combinations. Covered carriers start with one of the eight yaw angles; the carousel also has one of the three child angles. All dimensions below are in meters. Random draws are independent within their named stream unless the rule explicitly couples them.

A.2 A: ARTICULATED MARKER TRACKING

Scene. A blue link of length $l_1 \sim U(0.20, 0.25)$ lies along root Lx. A gold link of length $l_2 \sim U(0.12, 0.16)$ attaches at its endpoint, with a revolute joint about root Ly. A red sphere marks the distal endpoint. The purple joint cue identifies the hinge. The assembly is kinematically controlled; gravity does not make it fall. At relative angle q, the tip is

$$p_{\text{tip}} = h + R\big((l_1, 0, 0)^\top + R_y(q)(l_2, 0, 0)^\top\big).$$

Program and query. Each of T action slots is a child increment with probability 0.22, sampled from $\{-60, -30, 30, 60\}$ degrees. Otherwise, it is a root rotation about a sampled world/local axis and an angle from $\{-90, -60, -30, 30, 60, 90\}$. All intermediate child angles must remain within ± 120 degrees. The query asks for the final tip quadrant. Both $|p_x|$ and $|p_y|$ must exceed 0.02 m.

Twin. Keep the geometry, initial child angle, program, and question fixed. Enumerate other admissible root orientations in a seed-determined order until one gives the second required semantic answer. Reject the candidate root if none does. Both variants must pass independent projection and visibility checks.

A.3 B: HIDDEN IDENTITY IN A ROTATING CARRIER

Scene. Four compartments have local token centers $(r, r, 0.052)$, $(-r, r, 0.052)$, $(-r, -r, 0.052)$, and $(r, -r, 0.052)$, where $r \sim U(0.13, 0.15)$. The 0.052 m height is part of the rigid transform; it must not be discarded when the carrier tilts. The identities are a red sphere, blue cube, yellow cylinder, and green ellipsoid, assigned by a random permutation. Tokens are fixed to their compartments. They never slide, exchange places, or fall during rotations.

Encoding and program. Two initial views reveal all tokens; the third shows opaque lids closed without displacement. Each action rotates the root about a world/local axis by one of $\{-90, 90, 180\}$ degrees. The final configuration must place exactly one token center in every quadrant, with the 2 cm margins satisfied. The query identifies one fixed world quadrant and asks which token occupies it.

Twin and closed-state control. Exchange the nominal answer identity with the desired second identity. All positions and actions remain fixed. The revealing images change, but the closed images must have identical bytes. The renderer omits identity-dependent geometry in closed views so shadows or reflections cannot reveal a covered identity. This is a rendering rule for the opaque compartments, not an alteration of the private token state.

A.4 C: SUPPORT RELEASE WITH DISCLOSED MECHANICS

Scene. A uniform beam has half-length $b \sim U(0.25, 0.29)$, half-width 0.075, and half-thickness 0.014. A red cube has half-size $s \sim U(0.033, 0.040)$ and local center offset of magnitude $U(0.11, 0.18)$ with random sign. A purple axle marks an ideal hinge about beam Ly, initially at local offset in $\{-0.015, -0.010, 0, 0.010, 0.015\}$. Temporary clamps hold the beam level and the block at rest. All solids have density 500 kg/m^3. The rule card supplies the sampled beam/block dimensions and masses numerically. Block and pivot coordinates must be read from the images in the primary condition.

Program. The first $T - 1$ actions prepare the clamped fixture. Slots with zero-based index $i \bmod 3 = 1$ translate the pivot alone along current beam Lx by $\{-0.03, -0.02, 0.02, 0.03\}$ m. Other preparation slots yaw the whole clamped assembly about world $+z$ by $\{-90, 90, 180\}$ degrees. The pivot must remain inside the beam with a 0.04 m end margin. The final action releases both clamps. The hinge remains fixed; the beam rotates and the block moves under contact and gravity. The label is the block's resting floor quadrant. A one-action episode contains release only. Replacing release with a hold leaves the block clamped and yields no floor-receiver label, so that intervention is marked invalid in the action-sensitivity record rather than counted as an answer change.

Twin. Reflect the block's initial local offset, retaining the pivot, dimensions, yaw, and program. The desired semantic twin classes are opposite quadrants. Both outcomes must pass the numerical checks independently. The top-view projected block centers of the reflected alternatives must differ by more than 12 pixels.

Numerical contract. Table 2 fixes the engine, solver, and robustness gates; all physical labels are conditional on these settings.

Parameter	Value or acceptance rule
Engine and integrator	MuJoCo 3.13.0; `implicitfast`
Nominal/smaller steps	0.0005 / 0.00025 / 0.000125 s
Solver	Newton; 100 iterations; tolerance 10^{-10}; elliptic cones
Gravity and density	$(0, 0, -9.81)$ m/s^2; 500 kg/m^3

Contact friction	Sliding 0.5; torsional 0.005 m; rolling 0.0001 m; `condim`=3
Contact response	`solref`=`(0.0015,1)`; `solimp`=`(0.95,0.99,0.001,0.5,2)`
Hinge	Limits $[-50, 50]$ degrees; damping 0.04 N m s/rad
Rest	Block linear speed < 0.01 m/s; block and hinge angular speeds < 0.05 rad/s; block center below 0.10 m; sustained 0.5 s
Maximum duration	5 s after release
Floor contact	Normal force > 0.05 N for at least 0.02 s
Penetration	At most 0.002 m in every accepted run
Seeded perturbations	16 draws: translation per axis $U(-0.001, 0.001)$ m; rotation per axis $U(-0.5, 0.5)$ degrees; sliding friction/contact time constant scale $U(0.95, 1.05)$
Image-position corners	Block local offset and pivot local offset independently shifted by $\pm\epsilon$, $\epsilon = 2(1.5/768) = 0.00390625$ m; all four corners simulated

Table 2: Physical simulation and acceptance parameters.

Physical robustness tests explained. We rerun each accepted Task C release with time steps of 0.25 and 0.125 ms, compared with the nominal 0.5 ms. These smaller intervals between simulator updates test whether numerical resolution changes the final quadrant. We also run 16 trials using a fixed random seed for reproducibility. Each trial changes the block's initial position by up to 1 mm per world axis and its orientation by up to 0.5° per local axis, and independently scales sliding friction and the contact-response time constant by factors between 0.95 and 1.05. Sixteen is the protocol's chosen number of trials, not a statistical guarantee that every possible perturbation preserves the answer.

The position-uncertainty check addresses limited precision in reading positions from an image. Two pixels in the overhead view correspond to $\epsilon = 2(1.5/768)\,\mathrm{m} \approx 3.91$ mm. We shift the block's and pivot's positions along the beam independently by $-\epsilon$ or $+\epsilon$, testing all four combinations: $(-,-)$, $(-,+)$, $(+,-)$, and $(+,+)$. These are the "four corners" of the assumed position-error range, not corners of the objects. The two-pixel range is a declared tolerance, not a measured human perception error. We accept the scene only if all 22 additional simulations give the same final quadrant as the nominal run. These finite tests do not guarantee stability for every intermediate position or all possible physical parameters.

With beam mass m_b and block mass m_c, the initial combined center of mass along the beam is $x_{\mathrm{COM}} = m_c x_c/(m_b + m_c)$. We require

$$|x_{\mathrm{COM}} - x_p| - \epsilon\left(1 + \frac{m_c}{m_b + m_c}\right) > 0.$$

This is a torque-sign visibility margin, not the final answer oracle. The four position-corner outcomes are still simulated. Smaller-step and nuisance runs must agree with the nominal label and independently meet the rest/contact gates. The physical label remains conditional on the disclosed engine and parameters. With `condim`=3, only normal and tangential sliding contacts are active; the listed torsional and rolling parameter slots do not add rotational friction constraints. Floor tiles are visual quadrant markers; the physical floor is a flat plane.

A snapshot stores `mjSTATE_INTEGRATION`, including solver-related state, rather than only poses. The replay audit reconstructs an uninterrupted prefix from initialization, compares its state with the saved checkpoint at 0.5 s, and advances the uninterrupted and restored trajectories for 100 additional steps. The implementation requires prefix integration-state discrepancy below 10^{-10} and position-coordinate discrepancy below 10^{-12}. Visualization records use a separate restored state so rendering a trajectory cannot change the running simulation.

A.5 D: INDEPENDENT COUNTERFACTUAL BRANCHES

D uses A's mechanism and initial-state distribution. Sample a shared program of length $T-1$, then one root rotation for each branch, independently sampled from world/local axes and $\{-90, -60, 60, 90\}$ degrees. Execute the shared prefix once to obtain S_{T-1}, copy that state, and run each branch from its own copy. Let $b_A = \mathbf{1}[p_x^A > 0]$ and $b_B = \mathbf{1}[p_x^B > 0]$. The semantic classes are $(1,0)$, $(0,1)$, $(1,1)$, and $(0,0)$, presented as A only, B only, both, and neither. Both final positions require a 2 cm plane margin. The root-orientation twin changes this joint truth value. Sensitivity records for D cover the shared prefix, rather than counting the two branch suffixes as a single sequential program. Core branch suffixes have length one; the separate branch-consistency bank supplies longer branch relations.

A.6 E: COVERED CAROUSEL WITH ROOT AND JOINT UPDATES

E uses B's four-token geometry on a child carousel rotating about root Lz. Root arrows stay on the bearing frame, while a purple carousel-x cue turns with the child. A world token position is $h + RR_z(q)p_i$, retaining the token's nonzero local height. Even zero-based action slots increment the carousel angle by $\{-90, -60, 60, 90\}$ degrees; odd slots rotate the root about a world/local axis by $\{-90, 90, 180\}$. Both mechanisms must occur when $T \geq 2$. At $T \geq 4$, at least four single-action replacements by duration-matched holds must yield valid different answers:

$$D_{\text{sens}} = \sum_{t=1}^{T} \mathbf{1}[G_q(F(a_{1:t-1}, \text{hold}_t, a_{t+1:T}; S_0)) \neq y \text{ and the altered question remains valid}].$$

This requirement prevents accepting programs whose answer depends on very few individual slots. It does not imply that the program needs D_{sens} reasoning steps, or that its operations cannot be algebraically simplified. The hidden-identity twin and visibility conditions follow B.

A.7 STEP-BY-STEP FAMILY GENERATION

1. Choose a split, task, horizon, and root index. Hash the generator version, master seed, split, task, horizon, and index to obtain a private root seed. Use a separate two-root schedule to assign complementary semantic quotas and a common random option permutation.
2. For attempt j, derive an attempt seed. Sample mechanism dimensions, initial pose, identity assignment, camera, action program, and target query from named random streams. No model response enters this process.
3. Execute the nominal candidate. Reject illegal joint paths, insufficient workspace clearance, undefined receiver outcomes, final quadrant collisions, query-boundary cases, and candidates that miss the first semantic quota.
4. Construct the task-specific twin. Re-execute it and require the second semantic quota. Reject the whole root if the twin cannot be constructed.
5. For each variant, enumerate the declared initial-state alternatives at fixed geometry, check projected labeled landmarks, and render segmentation masks for target/identity visibility. For C, check reflected-position separation and the additional position/torque robustness contract.
6. Verify the kinematic trace with independent quaternion composition and MuJoCo forward kinematics. Record duration-matched single-action sensitivity. Apply E's minimum-sensitivity and mechanism-presence gates.
7. For each C variant, run the two smaller steps, 16 nuisance perturbations, four image-position corners, and snapshot continuation audit. Require all labels and acceptance criteria to agree. For B/E, require identical closed RGB bytes across the twin pair.
8. Accept the root only if both variants pass every applicable check. Otherwise record the rejection reason and try the next seed, up to 500 attempts. An unfilled quota is a generation error; it is not silently removed from the denominator.
9. Write private typed episodes, traces, certificates, and rejection counts. Render public initial packets with opaque filenames. Produce observation conditions from the accepted trajectory without resampling the scene or program.

10. Audit counts, balance, paired answers, option consistency, file existence, label re-execution, D suffixes, and C replay. Seal source hashes, private-root hashes, public-packet hashes, image hashes, and generation configuration before main test inference.

Existing successful roots resume unchanged. A changed source hash or generation configuration requires a new dataset directory. The development and test master seeds differ. No private seed, root identifier, certificate, or answer field is included in the model request. Public here means visible to the model in a named condition; it does not authorize releasing privileged FINAL/S0 packets as an unlabeled hidden-test distribution.

A.8 Rendering and distribution shifts

Table 3 specifies which attribute each extension changes and which controls remain fixed. All inputs are 768 by 768 pixels, with orthographic footprint 1.5 m. In-distribution oblique azimuths are $45, 135, 225, 315$ degrees and elevations are 35 or 55 degrees. The overhead camera has azimuth 90 degrees. Axis cues, identity colors/shapes, question text, and mechanics are fixed conventions. A conservative swept-radius bound keeps the kinematic mechanisms above the floor and inside the workspace; connected-part overlap at a designated hinge is allowed.

Extension	Exact change and scope
Geometry	A/D: first link $U(0.13, 0.16)$, second link $U(0.21, 0.24)$; B/E: carrier radius $U(0.16, 0.175)$; C: beam half-length $U(0.30, 0.33)$. Topology and identity conventions remain the same.
Camera	Oblique azimuth in $\{0, 90, 180, 270\}$, elevation 45 degrees. The overhead anchor remains unchanged; this is not complete viewpoint extrapolation.
Appearance	Darker floor and changed neutral materials. Identity colors/shapes and physical parameters remain unchanged.
Program	Root rotations use ± 120 degrees; C pivot increments use ± 0.01 or ± 0.04 m. Child-increment distributions remain unchanged. This holds out primitive magnitudes, not all program motifs.
Horizon stress	The same grammar at $T = 12, 16$. Acceptance checks and query margins are unchanged; stress scores remain outside the core macro-average.

Table 3: Distribution-shift parameters and retained controls.

B Auxiliary diagnostic construction

Auxiliary accuracy and relation metrics average within source task/horizon cells and then give those cells equal weight. Here "source horizon" refers to the parent core episode; it need not equal the derived question's action count. Bootstrap draws retain the original balancing-group dependencies among selected roots. Unequal retained group sizes are handled by recomputing each cell's mean over the roots in that draw. The separately reported overall invalid rate pools the diagnostic's individual questions. For all-success or all-failure samples, percentile bootstrap intervals collapse to the observed boundary; these are not population error bounds.

B.1 Missing, irrelevant, and decisive evidence

Select 50 B roots at each core horizon without model scores and use one variant per root. Construct three packets, each with three image slots and neutral view captions. The missing-evidence packet repeats the initial closed oblique view. The irrelevant-evidence packet inserts a closed overhead view between closed oblique views. The decisive packet supplies the two revealing initial views and the closed start. All three retain the same action program and query. The rule card explicitly says every assignment is admissible unless the images rule it out.

Enumerating the 24 identity permutations gives four possible token answers for either closed-only packet and one for the decisive packet, conditional on the accepted geometry. Option E is "not

determined by the supplied evidence." The correct response is E for both closed-only forms and the original token letter for the decisive form. A constant-E system obtains $2/3$ question accuracy but zero all-three-correct accuracy. Report accuracy by form, appropriate abstention on missing/irrelevant views, false abstention after decisive evidence, and all-three-correct root accuracy. These packets assess sensitivity to missing identity evidence, not general uncertainty calibration. The model receives identical text across the three forms; condition names and file paths are evaluation metadata and are absent from requests. A stronger analytical reference than constant abstention is a visibility-gated guesser: assuming perfect detection of whether any token identities are revealed, it abstains on closed-only packets and guesses uniformly among four tokens on the revealing packet. Its expected question accuracy is $5/6$ and its expected all-three-correct accuracy is 25%. These are calculated expectations, not measurements of an implemented visual classifier. Exceeding constant abstention alone therefore does not establish successful action tracking.

B.2 Divergence, reconvergence, and future equivalence

Select A roots in a fixed random order. Retain a prefix of at most two actions. Sample different root rotations a and b whose intermediate tip quadrants differ. For reconvergent roots, construct branch programs (a, a^{-1}) and (b, b^{-1}) after the prefix. Require equality of the complete kinematic endpoint state, including orientation and child angle. Append the same world-z, $+60$-degree rotation to both and verify equal future answers. For divergent roots, use (a, a^{-1}) and (b, b) and require different endpoint answers. Reject boundary cases and roots for which the specified relations cannot be constructed.

Each retained root produces six independent-context questions: the two intermediate states, two endpoints, and two common-suffix endpoints. The bank contains 50 reconvergent and 50 divergent roots. The source-horizon counts are $(18, 29, 28, 25)$ for $T = (1, 2, 4, 8)$; the reconvergent subset has counts $(11, 15, 8, 16)$. These unequal counts make the macro-average different from a pooled fraction of 100 roots. Actual intermediate, endpoint, and suffix questions contain respectively 2–3, 3–4, and 4–5 actions. Report per-query accuracy, all-six-correct accuracy, intermediate separation, endpoint-relation consistency, and future agreement on reconvergent roots. Each relation requires both answers in the queried pair to be valid; an invalid answer makes that pair's relation score zero. Invalid answers at another phase do not alter this local relation score. Agreement without correctness is insufficient evidence of correct tracking: a constant answer agrees across all reconvergent pairs but fails the required intermediate separation. This diagnostic checks output relations across fresh contexts; it does not test persistent memory across multiple conversational branches. Independent uniform four-way answers have expected pair correctness $4^{-2} = 6.25\%$ and all-six correctness $4^{-6} \approx 0.0244\%$. The low chance rate makes an all-six score near zero insufficient by itself to distinguish performance from guessing.

B.3 Hold padding and action sensitivity

Select 100 E roots at $T = 8$ whose core programs already satisfy $D_{\mathrm{sens}} \geq 4$. On the same scene, sample a valid two-action program consisting of a carousel increment followed by a world-z root rotation, and require $D_{\mathrm{sens}} = 2$. Insert these two actions, in the same order, into two sampled positions of an eight-slot program; fill the other positions with holds. The short and padded programs must have equal final labels, and the padded program must retain sensitivity two. The three resulting questions share the initial scene and option order. The short/padded difference isolates hold padding for these kinematic programs. The high-sensitivity/padded comparison also changes the active action program and sometimes the correct answer, so it is an associational comparison rather than a controlled estimate of reasoning depth. These are hold distractors, not unrelated-object interventions.

C Inference records and reproducibility

C.1 Commands and artifact layout

The analysis artifact contains source, tests, recorded score vectors, and five worked development examples with action replays. Statistical audits reproduce the historical, full-core, original and guided

follow-up intervals, and voting controls. The statistical audits use only supplied score vectors, without images, private labels, or raw model outputs. Figure and CSV regeneration is checked against recorded hashes. This verifies the analysis of stored scores, not the original model outputs. The archives remain local and the repository is private. The full dataset has no permanent public host. Identifying repository links are omitted for anonymous review. Appendix K documents the full-core extension and its separate provenance checks. The generator is a Python package with pinned lock files. A minimal regeneration and evaluation sequence is:

```
uv venv .venv --python 3.12
uv pip install --python .venv/bin/python -r requirements-generator.lock
.venv/bin/mentis3d generate --output data/example --split pilot \
  --seed 42 --roots-per-cell 2 --horizons 1,2,4,8 --conditions OL
.venv/bin/mentis3d audit data/example
.venv/bin/python scripts/audit_release.py data/example results/example_audit.json
.venv/bin/python scripts/serve_model.py qwen3vl-4b
.venv/bin/mentis3d evaluate data/example/public results/example_run \
  qwen3vl-4b --conditions OL --max-tokens 2048 --concurrency 32 \
  --response-mode brief \
  --model-revision ebb281ec70b05090aa6165b016eac8ec08e71b17
.venv/bin/mentis3d analyze data/example results/example_run
.venv/bin/python scripts/export_results.py results/example_run
```

Model weights are downloaded separately at the revisions in `configs/models.json`; they are not included in the source repository. Run the model server in a separate terminal, using the pinned inference environment described in the repository README. The example uses a disclosed demonstration seed. Private test seeds and answer files are held in the evaluation workspace. Exact hidden-bank reconstruction requires those private seeds; the published generator permits independently seeded replications of the construction distribution.

Table 4 distinguishes public inputs, privileged states, and retained inference records; these boundaries determine which files a model may receive in each condition.

Artifact	Contents and information boundary
`data/*/public/`	Condition-specific model packets and images. Core release packaging includes OL only; privileged packets remain evaluator inputs.
`data/*/private/`	Typed episodes, root membership, seeds, semantic labels, traces, physical states, and certificates. Never passed to a primary model.
`generation_config.json`	Split quotas, source hashes, version, and master seed. Private for hidden test banks.
`seal.json`	Content hashes and counts, without exporting seeds or labels.
`results/*/run_manifest.json`	Model/revision, endpoint, conditions, temperature, output budget, concurrency, parser version, and evaluator source hash.
`results/*/raw/`	Request hashes, image hashes, raw model responses, timing, token use, retries, and completion metadata. Kept locally because final-state responses can reveal test labels.
`results/*/summary.json`	Aggregate scores, uncertainty, and anonymized root/cluster score vectors, with no per-item answer letters.
`runtime/serving/`	Exact vLLM startup arguments and GPU metadata.

Table 4: Artifact layout and information boundaries.

The evaluator builds bounded numbers of base64 image requests to avoid memory use proportional to the entire dataset. It records raw responses before analysis, resumes completed packets, and retries only infrastructure failures when explicitly requested. It never retries an incorrect or invalid model answer until it becomes correct. One-step equivalent observation conditions share a request hash and response. Deterministic decoding is a protocol setting, not a claim of bitwise invariance across all GPU kernels or batching layouts.

After scoring both response contracts, `scripts/compare_contracts.py` checks exact root identity and exports paired, anonymized score vectors. The public commands `make results` and `make audit-public-statistics` rebuild manuscript numbers and reproduce their bootstrap intervals without private labels, raw responses, or image files. The full local integrity audit additionally checks requested packet identities, condition counts, server-response presence, and zero unresolved transport errors.

C.2 Statistical estimands and invalid responses

For cell c with R_c roots, question accuracy is

$$A_c = \frac{1}{2R_c} \sum_{r \in c} \sum_{v=0}^{1} c_{r,v}, \qquad P_c = \frac{1}{R_c} \sum_{r \in c} c_{r,0} c_{r,1}.$$

The core summaries average A_c and P_c over 20 cells. Bootstrap draws sample balancing groups within each cell with replacement, retaining their four questions. Matched differences use the same sampled group indices for both conditions. The larger-model comparison preserves five such groups per cell (400 questions); the observation panel preserves ten (800 questions). The larger-model allocation was reduced from 800 to 400 questions after observing serving latency, without examining test accuracy, using the same nested hash ordering. The incomplete 800-question attempt is not reported as a complete experiment. Confidence intervals summarize variation over this procedural sample; they do not include uncertainty from model training, alternative prompt wording, a different physics engine, or the uncollected participant-performance measurements. The intervals are marginal descriptive intervals; they are not adjusted for the many model, condition, and task comparisons. Small differences in leaderboard order should not be interpreted as established model rankings. An empirical percentile interval can collapse when every sampled group has the same score, including at zero or one. Such an interval describes the observed resamples and does not establish zero uncertainty about the wider population.

A terminal tag is recognized by a case-insensitive syntactic parser. The only normalizations are letter case and an optional terminal period. A response with no recognized tag, multiple tags, or an explanation after the tag is invalid. Invalid outputs count as incorrect. A response that reaches the token cap is not repaired or extended after observing its correctness. If it contains a valid terminal tag, it is scored by the same rule; truncation is also reported separately. HTTP failures and server exceptions are infrastructure failures and must be resolved before a run is called complete. Their archived attempts are not substituted for model responses.

Default online vLLM serving does not guarantee identical generated text across batch contexts (vLLM contributors, 2026). We did not enable its beta batch-invariance mode. Fixed seeds and greedy decoding therefore do not imply bitwise inference reproduction. The saved responses support exact reconstruction of the reported scores; the intervals condition on those responses and do not estimate variance over repeated serving runs.

The rejection counts in Table 5 describe the filtered sampling distribution; semantic quotas account for most rejected attempts. Table 6 confirms execution checks across the core and extensions, while Table 8 reports the narrower finite visible-evidence audit. Table 7 shows substantial variation in single-action sensitivity, including cells that admit insensitive programs. These quantities describe acceptance and program dependence, not human difficulty.

C.3 Acceptance distribution and extension verification

The factorized state equations matched 12,320 recorded configurations across the core and extension banks, including both D endpoints and C pre-release states. This audit uses exact initial states to characterize the programs' algebraic structure. The factorization check ends before C release; simulator trajectories are verified separately. A separate privileged torque-sign rule agrees with 1,760 of 1,760 nominal C labels across these banks. This is a post-generation characterization, never an acceptance gate. It shows that the coarse quadrant labels in this dataset admit an analytical shortcut; agreement does not establish a general law for other contact mechanisms. The visible-evidence audit retains only axis-label anchors with at least 90% of their text ink present in that view. Alternative root rotations must differ by more than two pixels at a retained anchor. At the identified root rotation, alternative allowed child angles must yield at least 100 symmetrically discrepant target/identity-mask

First recorded rejection reason	Attempts
semantic quota	8870
nonunique gate assignment	3140
query boundary margin	1411
twin semantic quota	408
branch predicate boundary	305
composite sensitivity below four	162
joint limit	157
physical torque ambiguous at image precision	104
physics rest timeout	87
no pose twin at required class	64
visual witness too small or occluded	12
physics penetration tolerance	2

Table 5: Core candidate rejections. The generator records the first failed gate per attempt; these counts are not independent failure probabilities. Semantic quotas account for most rejections.

Bank	Labels	D suffixes	C replays	C extra runs	Max. mm
test	4000	1600	800	17600	1.964
ood_geometry	1000	400	200	4400	1.879
ood_camera	1000	400	200	4400	1.835
ood_appearance	1000	400	200	4400	1.793
ood_program	1000	400	200	4400	1.818
stress	800	320	160	3520	1.787

Table 6: Post-generation execution checks for core and extensions. Extra runs are retained per-question physics stability certificates; the other columns count independently re-executed checks. Maximum penetration refers to the recorded accepted trajectories.

Task	$T = 1$	$T = 2$	$T = 4$	$T = 8$
A	0.27 [0, 1]	0.59 [0, 2]	1.29 [0, 4]	2.40 [0, 6]
B	1.00 [1, 1]	1.22 [0, 2]	2.36 [0, 4]	5.11 [2, 8]
C	0.00 [0, 0]	1.00 [1, 1]	2.00 [2, 2]	5.63 [5, 6]
D	0.00 [0, 0]	0.24 [0, 1]	0.85 [0, 3]	2.03 [0, 7]
E	0.79 [0, 1]	1.75 [0, 2]	4.00 [4, 4]	5.96 [4, 8]

Table 7: Observed single-action hold sensitivity: mean [minimum, maximum] across each core cell. D records cover the shared prefix only; replacing C release with a hold yields no floor label and is excluded. E has an acceptance floor of four sensitive actions at $T = 4, 8$. Other cells can include zero-sensitivity queries. This is not a minimum reasoning-depth certificate.

Bank	Nonphysical questions	Joint alternatives	Flags
Core	3200	4800	0
ood_geometry	800	1200	0
ood_camera	800	1200	0
ood_appearance	800	1200	0
ood_program	800	1200	0
stress	640	960	0

Table 8: Additional post-generation visible-evidence checks. These checks do not modify the frozen banks or supply masks to a model.

Baseline	Core accuracy	Core pair	Matched accuracy	Matched pair
Uniform sample	24.83 [23.50, 26.15]	6.75	23.25 [19.75, 27.00]	4.50
Static state (privileged)	33.03 [31.40, 34.65]	25.65	31.75 [26.75, 36.75]	25.00
Complete-state executor	100.00 [100.00, 100.00]	100.00	100.00 [100.00, 100.00]	100.00

Table 9: Programmatic baselines on the full core and the same matched 400 questions used for model comparisons. Rates are percentages. The uniform row is one seeded draw; its theoretical expected accuracy is 25%. The other two rows receive privileged initial state.

Model group	Native representation	Visual positions
Five Qwen checkpoints	48×48 patches, merge factor 2	576
InternVL3.5-8B	4 tiles plus thumbnail, 448^2 each	1280
SmolVLM2-2.2B	16 tiles plus thumbnail, 384^2 each	1377

Table 10: Pinned processing of one square 768-pixel RGB input. Counts refer to projected visual embedding positions, excluding caption and delimiter tokens. Native tiling/upsampling changes representation and compute but adds no new source-image information. These are not equal visual-token budgets.

pixels outside a two-pixel Chebyshev tolerance in one view. Masks exclude text overlays. The check holds sampled geometry and identity assignment fixed and assumes correct glyph and semantic identification; it is not an OCR or human-readability test. The renderer label positions and the audit coordinate mapping are checked across all declared initial poses.

C.4 Programmatic controls

Table 9 contrasts a uniform sample, a privileged static-state rule, and the complete-state executor. The executor verifies label computation; its access to hidden state prevents comparing it as an ordinary visual model. The static baseline predicts the initial projected block quadrant for C and repeats the initial positive-x predicate for both D branches. Undefined initial boundary queries count as incorrect. These are controls for action dependence and scoring, not systems with the primary visual information budget. Table 10 shows that equal source-image dimensions do not imply equal visual-token budgets across architectures.

C.5 Pinned model and serving configurations

The open-model identifiers below are immutable checkpoint revisions. These serving configurations describe the original matched panel; Appendix K gives the later full-core 27B settings and the separate Astra alias. The 27B model uses official FP8 weights. Its serving epochs record an initial CPU-offload profile and the subsequent reserved-cache GPU-only profile; completed responses are preserved across the change. The other models use the recorded serving dtype. Thinking is disabled only where the supplied chat template implements that switch. Primary requests explicitly override temperature and the output cap. Other settings retain pinned checkpoint and server defaults. The installed vLLM sampling audit confirms that temperature zero disables top-k and nucleus filtering, with zero presence/frequency penalties and unit repetition penalty; none of these checkpoint generation files overrides those penalties. The exploratory card-sampling run explicitly overrides its listed settings. End-of-sequence token IDs remain checkpoint-specific. The live SmolVLM tokenization check expands 3 images from one fixed pilot packet to 4131 visual positions (1377 per image), matching the native processor audit. This check generates no answer.

Qwen3-VL-4B


```
{
  "repository": "Qwen/Qwen3-VL-4B-Instruct",
  "revision": "ebb281ec70b05090aa6165b016eac8ec08e71b17",
  "gpu_memory_utilization": 0.85,
  "max_num_seqs": 32,
  "max_model_len": 16384,
  "processor": {
    "max_pixels": 589824
```

```
  }
}
```


Qwen3-VL-8B


```
{
  "repository": "Qwen/Qwen3-VL-8B-Instruct",
  "revision": "0c351dd01ed87e9c1b53cbc748cba10e6187ff3b",
  "gpu_memory_utilization": 0.88,
  "max_num_seqs": 24,
  "max_model_len": 16384,
  "processor": {
    "max_pixels": 589824
  }
}
```


Qwen3.5-4B


```
{
  "repository": "Qwen/Qwen3.5-4B",
  "revision": "851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a",
  "gpu_memory_utilization": 0.85,
  "max_num_seqs": 32,
  "max_model_len": 16384,
  "processor": {
    "max_pixels": 589824
  }
}
```


Qwen3.5-9B


```
{
  "repository": "Qwen/Qwen3.5-9B",
  "revision": "c202236235762e1c871ad0ccb60c8ee5ba337b9a",
  "gpu_memory_utilization": 0.88,
  "max_num_seqs": 24,
  "max_model_len": 16384,
  "processor": {
    "max_pixels": 589824
  }
}
```


InternVL3.5-8B


```
{
  "repository": "OpenGVLab/InternVL3_5-8B",
  "revision": "9bb6a56ad9cc69db95e2d4eeb15a52bbcac4ef79",
  "gpu_memory_utilization": 0.88,
  "max_num_seqs": 16,
  "max_model_len": 16384,
  "processor": {
    "max_dynamic_patch": 4
  },
  "trust_remote_code": true
}
```


Qwen3.8-27B-FP8


```
{
  "repository": "Qwen/Qwen3.8-27B-FP8",
  "revision": "017b9c7af6b5689d5dd426a76e0bc077eb5ca20a",
  "gpu_memory_utilization": 0.97,
  "max_num_seqs": 4,
  "max_model_len": 8192,
  "max_num_batched_tokens": 512,
  "cpu_offload_gb": 0,
  "processor": {
    "max_pixels": 589824
  },
  "enforce_eager": false,
  "kv_cache_memory_bytes": 1073741824
}
```

Cap	Accuracy [95% CI]	Pair [95% CI]	Invalid	Truncated
2048 tokens	10.00 [7.00, 14.00]	4.00 [0.00, 8.00]	77.00	77.00
4096 tokens	12.00 [8.00, 16.00]	0.00 [0.00, 0.00]	79.00	78.00

Table 11: Qwen3.8-27B-FP8 output-budget sensitivity on all 100 eight-action questions of the matched panel. The instruction is unchanged. Rates are percentages.

Decoding	Accuracy [95% CI]	Pair [95% CI]	Invalid	Truncated
greedy	10.00 [7.00, 14.00]	4.00 [0.00, 8.00]	77.00	77.00
card sampling	16.00 [11.00, 21.00]	4.00 [0.00, 10.00]	70.00	70.00

Table 12: Qwen3.8-27B-FP8 on the same 100 eight-action questions with the same brief instruction and 2,048-token cap. Card sampling uses the published non-thinking sampling/penalty settings. Rates are percentages.

SmolVLM2-2.2B


```
{
  "repository": "HuggingFaceTB/SmolVLM2-2.2B-Instruct",
  "revision": "482adb537c021c86670beed01cd58990d01e72e4",
  "gpu_memory_utilization": 0.82,
  "max_num_seqs": 32,
  "max_model_len": 8192,
  "max_images": 3,
  "processor": {},
  "dtype": "bfloat16"
}
```


For InternVL, the four-tile image limit is supplied through the runtime `max_dynamic_patch` model-configuration override. Passing it as a generic processor keyword in vLLM 0.29.0 also forwards it to an incompatible video constructor. The override preserves the specified image representation without editing checkpoint files. Serving manifests record this routing and the local dynamic-module cache. A live-server tokenization check verifies 1,280 visual positions per square image.

C.6 Output-budget sensitivity

Table 11 varies the output cap on the fixed eight-action 27B subset. Longer fresh responses do not necessarily continue the earlier text, so the contrast measures a serving run with a changed cap. The paired 4,096-minus-2,048-token accuracy difference is 2.00 [-4.00, 8.00] percentage points. This exploratory check changes the cap for every selected question, independent of its earlier correctness.

Exact generated text agrees through the end of the shorter response in 0 of 100 pairs. The median shared prefix is 194 characters. Fresh greedy requests produced different text across serving/batch contexts; this comparison is not equivalent to extending every original response verbatim.

C.7 Model-card sampling sensitivity

After inspecting the completed 27B primary and direct-answer reports, we added one exploratory run on all 100 matched eight-action questions using the model card's non-thinking settings (Qwen Team, 2026b): temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5, repetition penalty 1, and minimum-p 0. The brief instruction, 2,048-token cap, and question set are retained. The run uses the fixed sampling seed 20260910; its intervals do not include variation over decoding seeds. No parameter search or accuracy-based item selection is performed. A live-server tokenization check confirmed that the primary `enable_thinking=False` setting closes the thinking block before generation. Ordinary verbose answer text can still exceed the requested length. Table 12 compares greedy decoding with the model-card sampling bundle. Its paired interval reaches zero, and one bundled sampling setting does not identify an optimal decoder or an isolated temperature effect. The paired card-sampling-minus-greedy difference is 6.00 [0.00, 12.00] percentage points. This is one fixed sampling seed and a bundled configuration change, not an isolated temperature effect.

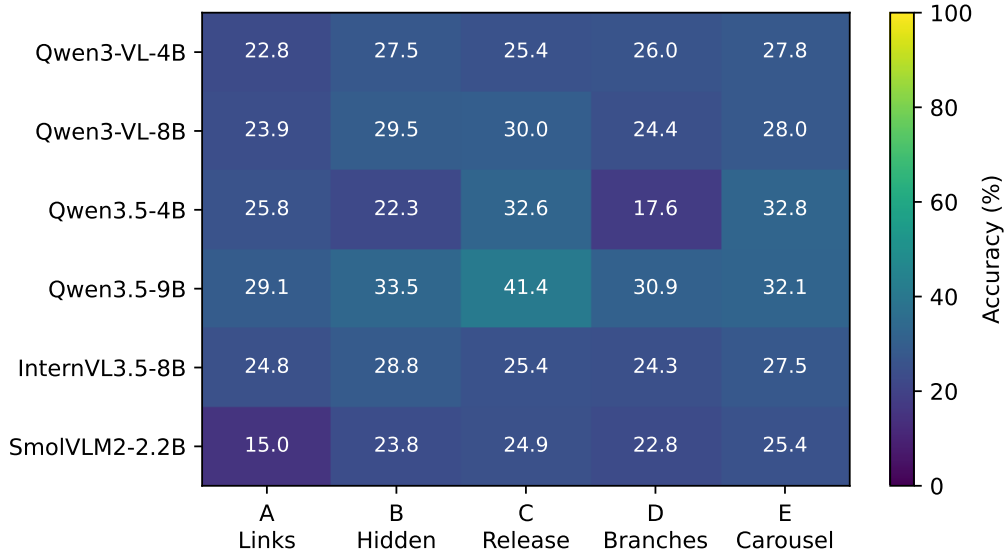


Figure 7: Primary accuracy by task family. Each model/task score averages four horizon cells containing 800 questions in total.

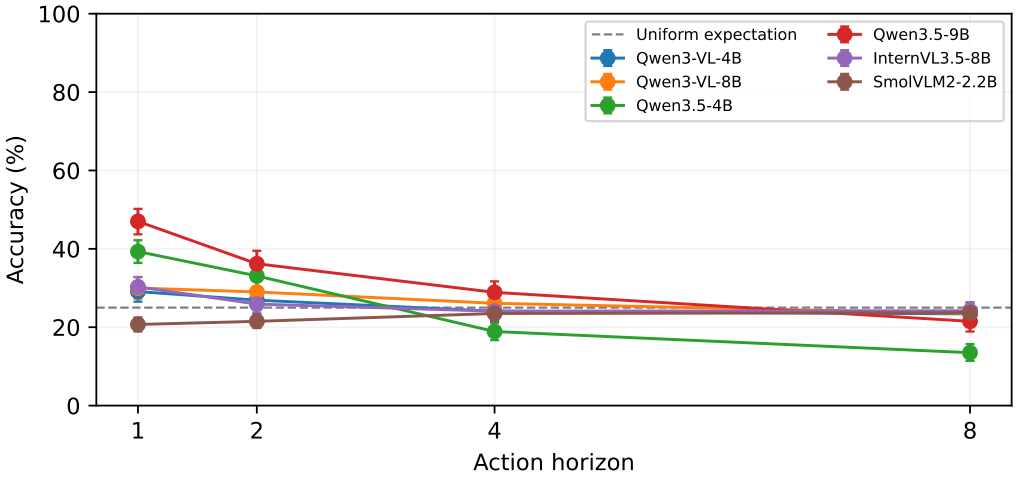


Figure 8: Primary accuracy by action horizon, averaging the five task families. Error bars show 95% balancing-group bootstrap intervals. Horizon cells contain independently generated roots. This comparison describes the sampled distributions; it is not a within-episode causal estimate of error accumulation. The separate hold-padding diagnostic controls terminal state.

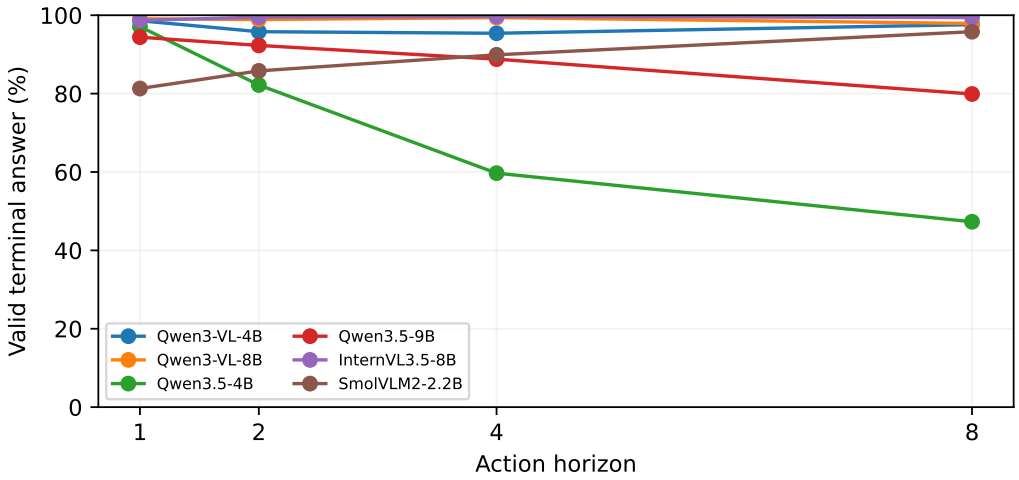


Figure 9: Observed valid-answer rates on the same full-core horizon cells, with 1,000 questions per model and horizon. A valid answer can be incorrect. These rates separate terminal-format compliance from the accuracy curves in Figure 8; they do not identify why a response is invalid.

Reading the detailed results. Table 13 retains the six original full-core runs; Table 54 adds the completed 27B and Astra runs. Table 14 instead holds the 400-question subset fixed across the seven open models. These panels have different denominators and, for the later 27B run, separately generated responses. The matched panel therefore supplies the appropriate reference for its intervention experiments.

Tables 15, 16, 17, 18, 19, and 20 decompose those six core runs into task/horizon cells. Reporting invalid responses beside accuracy distinguishes a missing usable answer from a valid wrong prediction. The cells also show why aggregate accuracy cannot establish a uniform failure pattern across tasks or horizons. Figures 7, 8, and 9 summarize these same runs by task, length, and response validity, respectively.

Information access. Tables 21 and 24 show strong terminal-coordinate readout alongside much lower initial-view accuracy. Tables 22 and 25 retain the task breakdown, while Tables 23 and 26 quantify paired changes. The repeated-image comparison is necessary because extra image slots can change responses even when they contain no updated state information. Neither exact-state nor image-feedback differences isolate a single internal reasoning component.

Shortcut and transfer controls. Tables 27 and 33 report the information-removal conditions. Their paired comparisons in Tables 28 and 34 use the same underlying questions, preventing a change in question membership from explaining the contrast. Tables 29 and 35 distinguish appropriate abstention from false abstention; all-three correctness requires both of these judgments and an answer when evidence is decisive. The branch tables (30, 36) report correctness beside agreement, since a constant answer can agree after reconvergence. The padding tables (31, 37) separate unchanged-endpoint padding from changes to the active program. Tables 32 and 38 give uncertainty for the independently generated shift and stress banks; these comparisons do not track the same scene under a controlled rerendering.

Output instructions and accounting. Tables 39, 40, 41, 42, 43, 44, and 45 compare brief and direct instructions on identical subsets. The direct contract changes both the instruction and output cap, so an accuracy difference cannot establish a benefit or cost of reasoning alone. Tables 46 and 47 show that even terminal readout changes under this contract. Table 48 records response lengths and termination counts, allowing readers to assess these scores alongside output completeness.

Model	Accuracy [95% CI]	Pair [95% CI]	Invalid
Qwen3-VL-4B	25.88 [24.55, 27.18]	6.15 [5.10, 7.20]	3.15
Qwen3-VL-8B	27.15 [25.80, 28.45]	6.90 [5.85, 7.95]	1.18
Qwen3.5-4B	26.20 [24.90, 27.53]	9.80 [8.60, 11.05]	28.43
Qwen3.5-9B	33.40 [31.93, 34.90]	14.80 [13.30, 16.25]	11.15
InternVL3.5-8B	26.13 [24.95, 27.30]	4.95 [4.05, 5.85]	0.68
SmolVLM2-2.2B	22.35 [21.48, 23.23]	2.15 [1.55, 2.75]	11.80

Table 13: Full-core aggregates: 4,000 questions per model. All rates are percentages.

Model	Accuracy [95% CI]	Pair [95% CI]	Invalid
Qwen3-VL-4B	28.00 [24.00, 32.00]	8.00 [5.00, 11.50]	3.00
Qwen3-VL-8B	28.50 [24.50, 32.50]	8.50 [5.00, 12.00]	1.00
Qwen3.5-4B	28.75 [24.50, 33.00]	13.00 [9.00, 17.50]	26.75
Qwen3.5-9B	34.00 [29.00, 38.75]	15.50 [10.50, 20.50]	10.50
InternVL3.5-8B	24.00 [21.00, 27.25]	3.50 [1.50, 6.00]	1.00
SmolVLM2-2.2B	23.50 [20.75, 26.50]	3.50 [1.50, 5.50]	11.75
Qwen3.8-27B-FP8	33.00 [29.25, 36.50]	22.00 [18.00, 26.00]	54.75

Table 14: Matched primary comparison: the same 400 questions for every model, under the brief-reasoning contract. All rates are percentages.

C.8 Detailed quantitative results

Intervals use the declared balancing-group bootstrap. All numbers in this appendix are generated from saved result files. For Qwen3.5-9B, all 400 prompt-only twin pairs have identical request hashes. Of these, 44 produce different response strings and 27 produce different parsed answers (including invalid outputs). Both twins are correct in 3 pairs. Nominally greedy serving is not bitwise repeatable in this recorded run; a nonzero observed prompt-only pair score does not contradict the exclusion for a strictly deterministic function of identical text. For InternVL3.5-8B, all 400 prompt-only twin pairs have identical request hashes. Of these, 0 produce different response strings and 0 produce different parsed answers (including invalid outputs). Both twins are correct in 0 pairs. All recorded twin responses are identical, consistent with the zero pair accuracy of a deterministic function of the shared text. The paired direct-minus-brief FINAL difference is -38.25 [-43.50, -32.75] percentage points. This post-hoc comparison uses existing responses and byte-identical question packets; both the instruction and output cap change. The paired direct-minus-brief FINAL difference is -56.25 [-60.25, -52.00] percentage points. This post-hoc comparison uses existing responses and byte-identical question packets; both the instruction and output cap change.

Table 48: Response diagnostics. Invalid, truncated, and over-150-word rates are percentages; tokens are mean generated tokens per logical packet, including reused packets. The word guide is not a scoring rejection criterion.

Model	Run/condition	n	Invalid	Trunc.	>150w.	Tokens
Qwen3-VL-4B	Core/OL	4000	3.15	2.28	8.40	225.2
Qwen3-VL-8B	Core/OL	4000	1.18	1.15	16.93	236.6
Qwen3.5-4B	Core/OL	4000	28.43	28.33	68.48	880.6
Qwen3.5-9B	Core/OL	4000	11.15	11.15	73.58	578.6
InternVL3.5-8B	Core/OL	4000	0.68	0.00	45.23	206.5
SmolVLM2-2.2B	Core/OL	4000	11.80	0.13	0.55	13.6
Qwen3-VL-4B	Match/OL	400	3.00	2.50	8.25	228.6
Qwen3-VL-8B	Match/OL	400	1.00	1.00	15.75	219.1
Qwen3.5-4B	Match/OL	400	26.75	26.50	69.50	845.2
Qwen3.5-9B	Match/OL	400	10.50	10.50	71.50	601.2
InternVL3.5-8B	Match/OL	400	1.00	0.00	45.50	207.0
SmolVLM2-2.2B	Match/OL	400	11.75	0.00	0.25	11.5
Qwen3.8-27B-FP8	Match/OL	400	54.75	54.75	98.50	1529.5
Qwen3-VL-4B	Direct/FINAL	400	0.00	0.00	0.00	4.2
Qwen3-VL-4B	Direct/OL	400	0.50	0.00	0.00	4.1
Qwen3-VL-8B	Direct/FINAL	400	0.00	0.00	0.00	4.0

Model	Run/condition	n	Invalid	Trunc.	>150w.	Tokens
Qwen3-VL-8B	Direct/OL	400	0.00	0.00	0.00	4.0
Qwen3.5-4B	Direct/FINAL	400	0.00	0.00	0.00	4.0
Qwen3.5-4B	Direct/OL	400	0.25	0.25	0.00	4.1
Qwen3.5-9B	Direct/FINAL	400	7.25	0.00	0.00	3.9
Qwen3.5-9B	Direct/OL	400	67.00	62.25	0.00	21.4
InternVL3.5-8B	Direct/FINAL	400	0.00	0.00	0.00	4.0
InternVL3.5-8B	Direct/OL	400	0.00	0.00	0.00	4.0
SmolVLM2-2.2B	Direct/FINAL	400	80.75	7.75	0.00	8.2
SmolVLM2-2.2B	Direct/OL	400	100.00	0.00	0.00	2.0
Qwen3.8-27B-FP8	Direct/FINAL	400	2.25	2.25	0.00	4.6
Qwen3.8-27B-FP8	Direct/OL	400	90.50	90.50	0.00	29.3
Qwen3.8-27B-FP8	Budget/OL	100	79.00	78.00	100.00	3683.2
Qwen3.8-27B-FP8	Card/OL	100	70.00	70.00	100.00	1803.1
Qwen3.5-9B	Obs/FB	800	9.50	9.50	73.38	537.1
Qwen3.5-9B	Obs/FINAL	800	0.25	0.25	12.25	184.9
Qwen3.5-9B	Obs/OL	800	10.25	10.25	71.50	574.8
Qwen3.5-9B	Obs/REP	800	10.63	10.63	74.00	548.4
Qwen3.5-9B	Obs/RESET	800	15.38	15.38	72.50	657.7
Qwen3.5-9B	Obs/S0	800	21.50	21.50	84.75	992.2
InternVL3.5-8B	Obs/FB	800	0.75	0.00	44.25	199.2
InternVL3.5-8B	Obs/FINAL	800	0.50	0.00	17.63	336.1
InternVL3.5-8B	Obs/OL	800	0.63	0.00	44.75	207.1
InternVL3.5-8B	Obs/REP	800	0.88	0.00	44.13	202.1
InternVL3.5-8B	Obs/RESET	800	1.75	0.00	48.88	207.6
InternVL3.5-8B	Obs/S0	800	2.50	0.00	82.25	275.9
Qwen3.5-9B	Short/encoding only	800	10.13	7.88	0.00	6.2
Qwen3.5-9B	Short/last action	800	36.75	28.13	0.00	11.8
Qwen3.5-9B	Short/prompt only	800	25.25	25.25	0.00	11.1
InternVL3.5-8B	Short/encoding only	800	0.00	0.00	0.00	4.0
InternVL3.5-8B	Short/last action	800	0.00	0.00	0.00	4.0
InternVL3.5-8B	Short/prompt only	800	0.00	0.00	0.00	4.0
Qwen3.5-9B	Branch/A end	100	12.00	11.00	57.00	528.1
Qwen3.5-9B	Branch/A mid	100	17.00	17.00	55.00	644.0
Qwen3.5-9B	Branch/A suffix	100	16.00	16.00	65.00	614.0
Qwen3.5-9B	Branch/B end	100	9.00	9.00	51.00	446.0
Qwen3.5-9B	Branch/B mid	100	15.00	14.00	58.00	571.2
Qwen3.5-9B	Branch/B suffix	100	23.00	23.00	67.00	746.2
Qwen3.5-9B	Depth/high	100	29.00	29.00	97.00	1184.6
Qwen3.5-9B	Depth/padded	100	7.00	7.00	86.00	664.8
Qwen3.5-9B	Depth/short	100	1.00	1.00	86.00	468.3
Qwen3.5-9B	Evid/decisive	200	3.50	3.50	88.50	452.9
Qwen3.5-9B	Evid/irrelevant	200	6.00	6.00	62.00	392.1
Qwen3.5-9B	Evid/missing	200	0.00	0.00	4.00	148.0
InternVL3.5-8B	Branch/A end	100	0.00	0.00	33.00	194.7
InternVL3.5-8B	Branch/A mid	100	0.00	0.00	16.00	180.6
InternVL3.5-8B	Branch/A suffix	100	0.00	0.00	60.00	215.4
InternVL3.5-8B	Branch/B end	100	0.00	0.00	24.00	189.0
InternVL3.5-8B	Branch/B mid	100	0.00	0.00	17.00	183.4
InternVL3.5-8B	Branch/B suffix	100	0.00	0.00	50.00	212.0
InternVL3.5-8B	Depth/high	100	0.00	0.00	63.00	229.3
InternVL3.5-8B	Depth/padded	100	0.00	0.00	64.00	221.4
InternVL3.5-8B	Depth/short	100	0.00	0.00	84.00	233.0
InternVL3.5-8B	Evid/decisive	200	2.50	0.00	52.50	233.4
InternVL3.5-8B	Evid/irrelevant	200	4.00	0.00	76.50	237.0
InternVL3.5-8B	Evid/missing	200	9.50	0.00	73.50	234.8
Qwen3.5-9B	Appear/OL	1000	19.20	19.20	89.60	771.6
Qwen3.5-9B	Camera/OL	1000	19.00	18.90	88.10	770.4
Qwen3.5-9B	Geom/OL	1000	16.00	15.90	87.70	737.8
Qwen3.5-9B	Magnitude/OL	1000	19.30	19.30	86.20	791.1
Qwen3.5-9B	Stress/OL	800	35.63	35.63	91.75	1141.6
InternVL3.5-8B	Appear/OL	1000	0.10	0.00	63.80	232.2

Model	Run/condition	n	Invalid	Trunc.	>150w.	Tokens
InternVL3.5-8B	Camera/OL	1000	0.20	0.00	65.00	234.6
InternVL3.5-8B	Geom/OL	1000	0.10	0.00	65.50	235.7
InternVL3.5-8B	Magnitude/OL	1000	0.30	0.00	60.80	218.8
InternVL3.5-8B	Stress/OL	800	0.25	0.00	54.63	209.4

Task	T	Accuracy	Pair	Invalid
A	1	25.50	4.00	2.50
A	2	21.00	5.00	11.50
A	4	20.00	5.00	5.50
A	8	24.50	5.00	1.00
B	1	34.50	15.00	1.50
B	2	27.50	9.00	4.50
B	4	28.50	9.00	3.00
B	8	19.50	6.00	3.50
C	1	24.50	2.00	0.00
C	2	27.00	3.00	0.50
C	4	21.50	1.00	5.00
C	8	28.50	4.00	2.00
D	1	26.50	3.00	1.50
D	2	27.50	5.00	2.00
D	4	27.00	6.00	0.00
D	8	23.00	5.00	0.00
E	1	34.50	13.00	1.50
E	2	31.50	11.00	2.50
E	4	23.50	5.00	9.50
E	8	21.50	7.00	5.50

Table 15: Qwen3-VL-4B: full-core task/horizon cells, 200 questions per cell.

Task	T	Accuracy	Pair	Invalid
A	1	21.50	3.00	2.00
A	2	25.50	3.00	3.00
A	4	23.00	4.00	0.00
A	8	25.50	6.00	0.00
B	1	36.50	21.00	1.00
B	2	29.50	12.00	1.00
B	4	29.00	9.00	3.00
B	8	23.00	5.00	5.50
C	1	29.00	3.00	0.50
C	2	36.50	7.00	0.50
C	4	30.50	6.00	0.00
C	8	24.00	4.00	1.00
D	1	26.00	3.00	0.50
D	2	24.50	7.00	0.00
D	4	24.00	2.00	0.00
D	8	23.00	5.00	0.00
E	1	37.00	20.00	0.50
E	2	29.00	8.00	1.00
E	4	24.00	3.00	0.00
E	8	22.00	7.00	4.00

Table 16: Qwen3-VL-8B: full-core task/horizon cells, 200 questions per cell.

Task	T	Accuracy	Pair	Invalid
A	1	27.50	10.00	7.00
A	2	32.50	15.00	8.00
A	4	25.00	8.00	16.00
A	8	18.00	2.00	35.50
B	1	43.50	22.00	1.50
B	2	35.00	22.00	13.00
B	4	7.00	1.00	75.00
B	8	3.50	0.00	89.00
C	1	41.50	10.00	0.50
C	2	34.00	8.00	2.50
C	4	26.50	6.00	4.00
C	8	28.50	9.00	9.50
D	1	33.00	6.00	5.00
D	2	22.00	9.00	54.00
D	4	6.50	0.00	76.50
D	8	9.00	1.00	55.50
E	1	51.00	34.00	0.50
E	2	42.00	22.00	11.50
E	4	29.50	9.00	30.00
E	8	8.50	2.00	74.00

Table 17: Qwen3.5-4B: full-core task/horizon cells, 200 questions per cell.

Task	T	Accuracy	Pair	Invalid
A	1	40.50	22.00	14.00
A	2	31.00	15.00	13.50
A	4	27.50	12.00	16.00
A	8	17.50	6.00	23.00
B	1	56.50	42.00	1.50
B	2	37.00	23.00	7.50
B	4	22.00	8.00	16.00
B	8	18.50	5.00	29.00
C	1	48.50	21.00	6.50
C	2	43.00	15.00	7.50
C	4	41.50	15.00	5.00
C	8	32.50	14.00	2.50
D	1	39.50	13.00	4.50
D	2	39.00	17.00	2.50
D	4	24.00	7.00	9.00
D	8	21.00	1.00	19.50
E	1	50.00	28.00	1.50
E	2	31.00	11.00	7.50
E	4	29.50	15.00	10.00
E	8	18.00	6.00	26.50

Table 18: Qwen3.5-9B: full-core task/horizon cells, 200 questions per cell.

Task	T	Accuracy	Pair	Invalid
A	1	28.50	6.00	0.00
A	2	23.50	3.00	0.00
A	4	23.50	3.00	0.00
A	8	23.50	4.00	0.00
B	1	36.00	15.00	0.00
B	2	26.00	8.00	0.00
B	4	24.00	10.00	0.00
B	8	29.00	9.00	0.00
C	1	27.00	1.00	3.50
C	2	29.50	1.00	0.00
C	4	23.00	1.00	0.50
C	8	22.00	0.00	2.00
D	1	24.00	2.00	3.00
D	2	24.00	4.00	2.50
D	4	29.00	2.00	0.50
D	8	20.00	0.00	1.00
E	1	36.00	13.00	0.00
E	2	26.50	8.00	0.00
E	4	21.00	4.00	0.50
E	8	26.50	5.00	0.00

Table 19: InternVL3.5-8B: full-core task/horizon cells, 200 questions per cell.

Task	T	Accuracy	Pair	Invalid
A	1	11.50	0.00	63.00
A	2	13.00	0.00	55.00
A	4	12.50	0.00	47.50
A	8	23.00	1.00	18.00
B	1	23.50	4.00	0.00
B	2	21.50	5.00	1.00
B	4	26.00	6.00	0.00
B	8	24.00	2.00	0.00
C	1	26.00	3.00	0.00
C	2	25.50	0.00	0.00
C	4	23.50	0.00	0.00
C	8	24.50	1.00	0.00
D	1	18.00	0.00	30.50
D	2	22.50	2.00	14.50
D	4	25.50	0.00	3.00
D	8	25.00	0.00	3.00
E	1	24.50	6.00	0.00
E	2	25.00	2.00	0.50
E	4	30.00	8.00	0.00
E	8	22.00	3.00	0.00

Table 20: SmolVLM2-2.2B: full-core task/horizon cells, 200 questions per cell.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FB	31.38 [28.13, 34.63]	12.75 [9.75, 15.75]	9.50
FINAL	95.38 [93.75, 96.88]	91.75 [89.00, 94.25]	0.25
OL	33.63 [30.13, 37.25]	15.75 [12.50, 19.25]	10.25
REP	34.25 [31.13, 37.38]	13.50 [10.50, 16.75]	10.63
RESET	32.13 [28.88, 35.38]	13.00 [10.00, 16.00]	15.38
S0	29.38 [26.25, 32.63]	13.25 [10.50, 16.25]	21.50

Table 21: Qwen3.5-9B: matched observation scores in percent.

Task	OL	FB	REP	S0	RESET	FINAL
A	26.88/11.88	26.88/9.38	30.63/9.38	25.00/30.63	26.25/20.00	100.00/0.00
B	34.38/15.00	25.63/10.63	30.00/12.50	37.50/28.75	29.38/20.63	89.38/0.63
C	44.38/3.13	36.88/6.88	44.38/3.13	24.38/18.13	40.63/10.00	99.38/0.00
D	32.50/11.88	30.00/10.63	31.25/13.13	31.88/15.00	28.75/18.75	99.38/0.00
E	30.00/9.38	37.50/10.00	35.00/15.00	28.13/15.00	35.63/7.50	88.75/0.63

Table 22: Qwen3.5-9B: family-level observation results. Each entry is accuracy/invalid rate in percent on the same 160 questions; family-level accuracy intervals are retained in the public summaries.

Comparison	Difference [95% CI]
FB minus OL	-2.25 [-5.50, 1.00]
FINAL minus OL	61.75 [57.88, 65.50]
REP minus OL	0.63 [-2.63, 3.88]
RESET minus OL	-1.50 [-5.13, 2.00]
S0 minus OL	-4.25 [-8.88, 0.25]
FB minus REP	-2.88 [-6.00, 0.25]
FB minus OL ($T > 1$)	-3.00 [-7.17, 1.33]
FB minus REP ($T > 1$)	-3.83 [-7.83, 0.33]

Table 23: Qwen3.5-9B: paired accuracy differences in percentage points.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FB	26.38 [23.88, 28.88]	4.50 [2.75, 6.50]	0.75
FINAL	98.75 [97.88, 99.50]	97.75 [96.25, 99.00]	0.50
OL	26.63 [24.13, 29.13]	5.00 [3.00, 7.00]	0.63
REP	27.38 [24.50, 30.25]	6.50 [4.25, 9.00]	0.88
RESET	27.25 [24.75, 29.75]	5.00 [3.00, 7.00]	1.75
S0	26.88 [23.88, 30.00]	9.00 [6.50, 11.75]	2.50

Table 24: InternVL3.5-8B: matched observation scores in percent.

Task	OL	FB	REP	S0	RESET	FINAL
A	26.88/0.00	28.13/0.00	29.38/0.63	26.25/0.00	28.13/0.00	98.75/0.00
B	27.50/0.00	25.63/0.00	28.13/0.00	29.38/0.00	25.00/0.00	99.38/0.00
C	26.25/2.50	30.00/2.50	29.38/2.50	24.38/0.00	28.13/3.75	100.00/0.00
D	26.25/0.63	26.25/1.25	26.25/1.25	21.25/12.50	30.00/5.00	97.50/2.50
E	26.25/0.00	21.88/0.00	23.75/0.00	33.13/0.00	25.00/0.00	98.13/0.00

Table 25: InternVL3.5-8B: family-level observation results. Each entry is accuracy/invalid rate in percent on the same 160 questions; family-level accuracy intervals are retained in the public summaries.

Comparison	Difference [95% CI]
FB minus OL	-0.25 [-2.88, 2.38]
FINAL minus OL	72.13 [69.62, 74.63]
REP minus OL	0.75 [-2.00, 3.50]
RESET minus OL	0.63 [-2.13, 3.50]
S0 minus OL	0.25 [-3.63, 4.13]
FB minus REP	-1.00 [-4.25, 2.25]
FB minus OL ($T > 1$)	-0.33 [-3.83, 3.17]
FB minus REP ($T > 1$)	-1.33 [-5.67, 2.83]

Table 26: InternVL3.5-8B: paired accuracy differences in percentage points.

Control	Accuracy [95% CI]	Pair [95% CI]
encoding_only	23.38 [20.88, 25.88]	5.75 [3.75, 8.00]
last_action	16.00 [14.00, 18.00]	3.50 [2.00, 5.25]
prompt_only	19.13 [17.50, 20.88]	0.75 [0.00, 1.50]

Table 27: Qwen3.5-9B: direct-answer shortcut controls, 800 questions per condition.

Control	Matched accuracy	Invalid	Control − OL
OL reference	11.00 [9.00, 13.00]	67.00	–
encoding_only	26.50 [23.25, 29.75]	9.25	15.50 [12.00, 19.00]
last_action	16.75 [14.25, 19.50]	39.25	5.75 [3.25, 8.25]
prompt_only	19.75 [18.00, 21.50]	25.00	8.75 [6.25, 11.25]

Table 28: Qwen3.5-9B: the same 400 questions under the same direct-answer instruction and 32-token cap. Differences are paired percentage points; accuracy and invalid rates are percentages. Differences include changes in format compliance.

Condition or metric	Estimate [95% CI]
EVIDENCE_decisive	33.00 [27.48, 38.78]
EVIDENCE_irrelevant	77.50 [72.07, 82.82]
EVIDENCE_missing	100.00 [100.00, 100.00]
all_correct	23.50 [18.05, 29.00]
appropriate_abstention	88.75 [86.01, 91.35]
false_abstention	9.50 [5.44, 13.89]
decisive_minus_missing	-67.00 [-72.75, -61.35]

Table 29: Qwen3.5-9B: evidence diagnostics; accuracy/agreement in percent and differences in percentage points.

Condition or metric	Estimate [95% CI]
BRANCH_A_end	42.31 [32.25, 52.87]
BRANCH_A_mid	27.59 [18.20, 37.09]
BRANCH_A_suffix	36.81 [27.79, 45.76]
BRANCH_B_end	33.44 [24.53, 43.19]
BRANCH_B_mid	34.69 [25.60, 43.97]
BRANCH_B_suffix	24.98 [16.41, 34.81]
all_correct	0.86 [0.00, 2.59]
mid_separation	51.54 [42.23, 60.94]
end_relation_consistency	59.17 [49.97, 68.04]
mid_both_correct	12.18 [6.10, 18.90]
end_both_correct	20.95 [13.47, 29.28]
suffix_both_correct	13.24 [6.87, 20.78]
relation_and_answers_correct	0.86 [0.00, 2.59]
future_agreement	29.07 [16.75, 42.14]

Table 30: Qwen3.5-9B: branches diagnostics; accuracy/agreement in percent and differences in percentage points.

Condition or metric	Estimate [95% CI]
DEPTH_high	26.00 [17.00, 35.00]
DEPTH_padded	33.00 [24.00, 42.00]
DEPTH_short	37.00 [29.00, 45.00]
all_correct	7.00 [3.00, 12.00]
padded_minus_short	-4.00 [-15.00, 8.00]
high_minus_padded	-7.00 [-20.00, 5.00]
padded_short_answer_agreement	36.00 [27.00, 45.00]

Table 31: Qwen3.5-9B: depth diagnostics; accuracy/agreement in percent and differences in percentage points.

Extension	Accuracy [95% CI]
Core reference ($T = 4, 8$)	25.20 [23.25, 27.10]
ood_geometry	27.90 [25.30, 30.60]
ood_camera	23.00 [20.60, 25.40]
ood_appearance	25.90 [23.20, 28.60]
ood_program	26.30 [23.60, 29.10]
stress	17.00 [14.63, 19.38]

Table 32: Qwen3.5-9B: extension intervals.

Control	Accuracy [95% CI]	Pair [95% CI]
encoding_only	24.38 [22.38, 26.38]	5.00 [3.25, 7.00]
last_action	23.50 [21.50, 25.50]	4.75 [3.00, 6.75]
prompt_only	25.50 [23.88, 27.13]	0.00 [0.00, 0.00]

Table 33: InternVL3.5-8B: direct-answer shortcut controls, 800 questions per condition.

Control	Matched accuracy	Invalid	Control − OL
OL reference	25.75 [22.50, 29.00]	0.00	–
encoding_only	26.25 [23.25, 29.50]	0.00	0.50 [-2.50, 3.50]
last_action	25.25 [22.25, 28.50]	0.00	-0.50 [-3.25, 2.25]
prompt_only	27.00 [24.75, 29.25]	0.00	1.25 [-2.00, 4.50]

Table 34: InternVL3.5-8B: the same 400 questions under the same direct-answer instruction and 32-token cap. Differences are paired percentage points; accuracy and invalid rates are percentages. Differences include changes in format compliance.

Condition or metric	Estimate [95% CI]
EVIDENCE_decisive	32.00 [25.65, 38.34]
EVIDENCE_irrelevant	24.00 [18.50, 29.71]
EVIDENCE_missing	25.50 [20.06, 31.20]
all_correct	2.00 [0.47, 4.05]
appropriate_abstention	24.75 [20.55, 29.07]
false_abstention	3.00 [0.98, 5.52]
decisive_minus_missing	6.50 [-2.08, 15.08]

Table 35: InternVL3.5-8B: evidence diagnostics; accuracy/agreement in percent and differences in percentage points.

Condition or metric	Estimate [95% CI]
BRANCH_A_end	29.36 [20.83, 38.28]
BRANCH_A_mid	25.04 [16.82, 34.24]
BRANCH_A_suffix	19.26 [12.03, 27.03]
BRANCH_B_end	24.63 [16.68, 33.23]
BRANCH_B_mid	20.22 [12.43, 28.27]
BRANCH_B_suffix	31.25 [22.75, 40.56]
all_correct	0.00 [0.00, 0.00]
mid_separation	48.44 [38.71, 58.09]
end_relation_consistency	56.69 [46.93, 65.96]
mid_both_correct	1.75 [0.00, 4.56]
end_both_correct	10.18 [4.65, 16.35]
suffix_both_correct	10.44 [5.15, 16.44]
relation_and_answers_correct	0.00 [0.00, 0.00]
future_agreement	50.51 [37.09, 64.39]

Table 36: InternVL3.5-8B: branches diagnostics; accuracy/agreement in percent and differences in percentage points.

Condition or metric	Estimate [95% CI]
DEPTH_high	25.00 [18.00, 32.00]
DEPTH_padded	16.00 [9.00, 23.00]
DEPTH_short	28.00 [20.00, 37.00]
all_correct	0.00 [0.00, 0.00]
padded_minus_short	-12.00 [-24.00, 0.00]
high_minus_padded	9.00 [-3.00, 20.00]
padded_short_answer_agreement	29.00 [20.00, 38.00]

Table 37: InternVL3.5-8B: depth diagnostics; accuracy/agreement in percent and differences in percentage points.

Extension	Accuracy [95% CI]
Core reference ($T = 4, 8$)	24.15 [22.55, 25.75]
ood_geometry	27.40 [24.80, 30.00]
ood_camera	25.60 [23.20, 27.90]
ood_appearance	24.20 [21.90, 26.50]
ood_program	25.30 [22.90, 27.60]
stress	24.13 [21.63, 26.63]

Table 38: InternVL3.5-8B: extension intervals.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	45.25 [41.00, 49.75]	23.00 [18.00, 28.50]	0.00
OL	25.25 [21.75, 28.75]	5.00 [2.50, 8.00]	0.50

Table 39: Qwen3-VL-4B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	49.50 [45.25, 54.00]	23.50 [18.00, 29.50]	0.00
OL	28.25 [24.75, 31.75]	6.50 [4.00, 9.50]	0.00

Table 40: Qwen3-VL-8B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	49.00 [45.25, 53.00]	26.00 [21.00, 31.50]	0.00
OL	27.25 [24.25, 30.25]	6.00 [3.50, 8.50]	0.25

Table 41: Qwen3.5-4B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	55.75 [51.00, 60.75]	37.00 [31.00, 43.50]	7.25
OL	11.00 [9.00, 13.00]	2.00 [0.50, 3.50]	67.00

Table 42: Qwen3.5-9B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	42.75 [38.75, 47.00]	16.00 [11.00, 21.50]	0.00
OL	25.75 [22.50, 29.00]	5.00 [2.50, 7.50]	0.00

Table 43: InternVL3.5-8B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	4.75 [4.00, 5.50]	0.00 [0.00, 0.00]	80.75
OL	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	100.00

Table 44: SmolVLM2-2.2B: direct-answer matched comparison, 400 questions per condition.

Condition	Accuracy [95% CI]	Pair [95% CI]	Invalid
FINAL	90.50 [87.75, 93.25]	85.00 [80.50, 89.00]	2.25
OL	2.75 [1.50, 4.25]	0.50 [0.00, 1.50]	90.50

Table 45: Qwen3.8-27B-FP8: direct-answer matched comparison, 400 questions per condition.

Contract	Accuracy [95% CI]	Pair [95% CI]	Invalid
brief	94.00 [91.50, 96.50]	89.50 [85.00, 93.50]	0.50
direct	55.75 [51.00, 60.75]	37.00 [31.00, 43.50]	7.25

Table 46: Qwen3.5-9B: FINAL readout on the same 400 preselected questions, nested within the 800-question observation panel. The brief contract allows 2,048 tokens and the direct contract 32. Rates are percentages.

Contract	Accuracy [95% CI]	Pair [95% CI]	Invalid
brief	99.00 [98.00, 99.75]	98.00 [96.00, 99.50]	0.50
direct	42.75 [38.75, 47.00]	16.00 [11.00, 21.50]	0.00

Table 47: InternVL3.5-8B: FINAL readout on the same 400 preselected questions, nested within the 800-question observation panel. The brief contract allows 2,048 tokens and the direct contract 32. Rates are percentages.

D HUMAN DATASET VALIDATION AND PARTICIPANT-REVIEW PROTOCOL

D.1 COMPLETED DATASET CORRECTNESS REVIEW

Two human reviewers checked approximately 10% of the benchmark dataset and certified the correctness of the reviewed subset. We report this as a qualitative correctness check. We do not report human answer accuracy, completion time, inter-rater agreement, or a readability acceptance rate. The separate review interface and automated UI tests verify software behavior; no response files from its instrumented participant study have been imported.

D.2 ALLOCATION AND PARTICIPANT PROCEDURE

Table 49 separates the completed dataset review from the instrumented participant-study allocation. The review bank selects ten complete roots per task/horizon cell, giving 200 roots and 400 questions. Five worked examples use other development roots. Thirty participant numbers index 40-question forms. The allocation assigns three ratings to each question if all forms are completed by different people, and never assigns both variants of one root to the same person. Reusing a participant number or assigning several numbers to one person does not create independent ratings. The organizer must enforce one person per number.

Before recruitment, obtain the applicable institutional review and consent determination. Recruit adult readers who understand elementary Cartesian coordinates after instruction. Record prior familiarity. The tutorial explains world and local axes, the right-hand rule, hinge directions, token attachment, and independent branch starts. Participants can apply the worked-example actions before selecting an answer. This common preparation must be reported when interpreting subsequent accuracy.

Each question presents all initial views, the complete rules, the action list, and four options. Participants choose an option or explicitly skip with a reason, report confidence on a five-point scale, then lock the response. There is no response deadline. Active solving time begins after initial images load and pauses when the page is hidden or the participant pauses the session. Elapsed time retains interruptions; feedback time is recorded separately. After locking, the answer and its solving time cannot be edited through the interface.

D.3 APPLYING ACTIONS FOR VISUAL CROSS-CHECKING

`Apply next action` is available before answering, with controls to return to a previous state, reset to the initial state, or select any action boundary. Each state has side-oblique, overhead, and opposite-side views. Tokens are revealed for explanation, even when the scored model input keeps them covered. Coordinate tables identify each token's world position and quadrant. The final state displays the verified answer. These views therefore provide substantial assistance and must be distinguished from the primary visual prediction task.

The replay contains 2,006 states and 6,018 images across the 400 study questions and five examples. States come from the existing development execution records. Side cameras use the original azimuth and its opposite, each at 20 degrees above horizontal; the overhead camera stays in a fixed world orientation. Both D branches restart from the shared-prefix endpoint. A C release view uses the stored accepted resting snapshot, without claiming to animate or independently rerun the falling trajectory. Image integrity and action-boundary coverage were checked for every asset. Separate quaternion checks cover all 162 displayed branch endpoints.

Opening a replay before answer lock records `replay.before_answer=true` under `interactive_action_replay_v2`. The record retains selected state keys, image enlargements, loading issues, and exposure to a previous question's replay. Such answers are excluded from unaided accuracy and clean timing summaries. Post-answer-only inspection preserves the current locked answer but may influence later questions. Reports therefore separate assistance on the current item from previous replay exposure. Legacy records without exposure fields cannot establish absence of assistance outside the instrumented interface. The static page embeds explanatory states; it is a review tool, not a secure test-delivery system.

D.4 Feedback and durable records

Participants rate visibility of token identity, world axes, local axes, joints, and targets; clarity of rotation direction, reference frame, action sequence, mechanical rules, and options; sufficiency of the original information; and five-point difficulty. They can mark absent features as not applicable. Issue tags identify occlusion, unreadable labels, ambiguous directions, missing information, and other problems. A flagged issue requires an explanation; reviewers are asked to name the action and view. Suggested fixes and external aids are recorded separately. Ratings concern the original question, even when replay assists the subsequent diagnosis.

The Jupyter form stores browser progress and exports cumulative JSON snapshots. Each export contains a random session identifier, assigned participant number, bank hash, question order, completed responses, timing, ratings, and replay exposure. The optional local-server mode saves to SQLite. The JSON importer checks the bank, allocation, response schema, and ordered completed prefix. It merges additional responses without overwriting prior answers or double-counting repeated exports. Accepted originals are archived privately by content hash. Structural validation does not authenticate participant identity or browser measurements. Feedback must not include names or other identifying information.

D.5 Analysis and adjudication

The implemented reports provide descriptive accuracy including skips, answered-only accuracy, time medians and quartiles, and feedback distributions by task and horizon. They separate visually assisted responses, post-answer replay, and records without replay. Clean timing also excludes external aids, technical problems, pauses, and hidden-tab events. A separate summary restricts to answers without current assistance or recorded earlier replay exposure. Participant/item dependence precludes treating all responses as independent; the current review summaries are descriptive. Because replay use is self-selected, conditional unaided accuracy does not estimate accuracy for a population forced to solve without assistance, and assisted-minus-unaided differences are not causal effects. A confirmatory unaided baseline requires a separately specified cohort and information-access protocol before recruitment.

At least three distinct assigned raters are required before item-level triage. At least two reports of a visibility, clarity, technical, or information problem trigger inspection. Low accuracy alone triggers difficulty review, not automatic exclusion. The adjudication log records a retain, revise, or future-version exclusion decision with a reviewer identifier, rationale, source-report hash, and affected balancing group. Exclusion proposals require at least three ratings. Changes to construction rules require a new dataset version and independent reevaluation; the present bank and model scores remain fixed.

Human-review artifact or measurement	Current status
Dataset correctness reviewers	2
Dataset coverage	Approximately 10%
Reviewed subset	Certified correct
Allocated development questions	400
Allocated participant forms	30
Questions per form	40
Instrumented participant sessions	0
Imported study response files	0
Human accuracy and completion time	Not measured

Table 49: Completed dataset validation and separate participant-study status. Two humans checked approximately 10% of the benchmark dataset and certified the reviewed subset as correct. The instrumented study has no recorded participant sessions; its allocated forms are not performance observations.

Reproducing the interface. After preparing the development study bank and worked examples, run:

```
.venv/bin/python scripts/build_human_replay.py
```

```
.venv/bin/python scripts/build_jupyter_study.py
.venv/bin/python scripts/audit_human_replay.py
.venv/bin/python scripts/import_human_json.py
.venv/bin/python scripts/analyze_human_feedback.py
```

Open `study/jupyter.html` in Jupyter's HTML viewer and enable Trust HTML. Response JSON files are uploaded to `data/human_study/incoming/`. Generated replay images and private feedback remain outside the source archive.

E DATASET DOCUMENTATION AND ACCESS

Table 50 records provenance and information access, including the distinction between model-visible packets and private labels.

Field	Description
Purpose	Controlled evaluation of state tracking under disclosed 3D actions, with paired questions and information interventions. Intended users include multimodal-evaluation and spatial-reasoning researchers.
Provenance	Procedural primitive scenes and templated instructions; the renderer supplies images and executable queries supply labels. Generation uses no scraped images, photographed people, or human answer annotations.
Unit and composition	Each root contributes two questions with the same language and different semantic answers. Two roots form a four-question balancing group. Table 1 lists the main banks; auxiliary evidence, branch, and padding diagnostics retain separate denominators.
Fields	Public packets include opaque item IDs, task, horizon, instructions, image paths and captions, action text, question, and options. Private episodes additionally store construction identity, parameters, initial state, actions, semantic label, option permutation, trace, and certificates.
Sampling	Named seeded streams sample scene and program attributes. Acceptance depends on answer quotas, visual evidence, geometric constraints, and physical stability. The core acceptance rate is 11.96%; the distribution is conditional on these filters.
Splits	Pilot, development, and test use distinct construction seeds. Both variants and all information conditions of a root remain together. The participant-study allocation uses development roots; test model responses never determine acceptance.
Verification	Executable query functions generate labels; independent rotations, simulator landmarks, physical perturbations, and visibility checks test specified failure modes. Two human reviewers also checked approximately 10% of the benchmark dataset and certified the reviewed subset as correct. These checks do not establish unaided human performance or universal physical validity.
Missing values	Invalid model output counts as incorrect under the specified parser. Infrastructure failures leave a run incomplete until repaired. Unevaluated model/bank combinations and uncollected human measurements are explicitly marked.
Access	The private repository holds source and summaries; local archives hold full images, episodes, and raw responses. An anonymous analysis archive contains source and score vectors. No permanent public host or dataset identifier exists.
Licensing	The dataset card assigns Apache License 2.0 to code and generated assets. Model weights and the ICLR template retain their own licenses; the benchmark does not redistribute model weights.
Maintenance	Content hashes identify source, banks, requests, and reports. An append-only adjudication log records decisions for complete balancing groups. Revisions receive new versions and preserve prior results.
Human data	Two reviewers certified approximately 10% of the benchmark dataset as correct. No participant-level accuracy or timing data are included. The review interface uses assigned participant numbers, private storage, and explicit assistance records. Participant recruitment requires a separate consent and institutional-review protocol.

Limits of use	The benchmark does not represent unrestricted everyday 3D scenes, general contact planning, or safety-critical physics. Scores are behavioral measurements under stated inference contracts and do not establish mental imagery, consciousness, or a particular latent mechanism.

Table 50: Dataset provenance, access, and use constraints.

F COMPARISON WITH RELATED EVALUATION SETTINGS

Table 51 compares the questions these benchmarks ask and the evidence they supply. It is a scope comparison, not a claim that every unlisted feature is absent. Scores across these datasets have different tasks, models, and inference contracts and are not directly comparable.

Benchmark	Established evaluation setting	Relation to this benchmark
MentisOculi (Zeller et al., 2026)	Procedural multistep visual reasoning and analyses of intermediate visualizations.	Motivates the procedural setting and visual-aid controls. Our contribution specifies paired 3D mechanisms, observation checks, and matched state-access conditions.
STARE (Li et al., 2026)	Geometric transformations, spatial puzzles, and visual simulation aids, with human and model evaluation.	Establishes that intermediate visual aids and 3D transformations are prior ideas. We study prescribed mechanism actions with paired labels and numerical verification.
MindCube (Wang et al., 2026)	Limited-view reasoning about positions, perspective, and hypothetical motion; cognitive-map scaffolding.	Motivates explicit spatial state and unseen-space reasoning. Our exact-state inputs are diagnostic conditions on fixed questions, not a trained mapping method.
CLEVRER (Yi et al., 2020)	Descriptive, explanatory, predictive, and counterfactual questions about physical-event videos.	Establishes controlled causal and counterfactual physical reasoning. Our main input is initial views plus an action program, with one restricted support-release grammar.
ReVSI (Zhang et al., 2026b)	Visual spatial evaluation that checks the information actually supplied in the input.	Motivates separating answerability from correctness. Our acceptance checks operate on declared construction families and retain their finite-state limitations.
MENTISOCULI-3D	Five question families over three mechanism grammars, paired initial states, and matched information interventions.	Supplies verified traces, group-preserving splits and intervals, response-validity reporting, and a development review tool with recorded assistance. Two humans certified approximately 10% of the benchmark dataset as correct; human performance remains unmeasured.

Table 51: Closest conceptual overlaps and the scope of the implemented contribution.

F.1 WHAT PAIRED ACCURACY DOES AND DOES NOT SHOW

Let a pair have identical textual input x and distinct answer letters $y_0 \neq y_1$. A deterministic text-only predictor $f(x)$ gives the same letter on both variants, so

$$\mathbf{1}\{f(x) = y_0\}\mathbf{1}\{f(x) = y_1\} = 0.$$

Its average question accuracy on the pair is at most $1/2$. This guarantee concerns predictors that see only the shared text; opaque item identifiers and other metadata are excluded from the request.

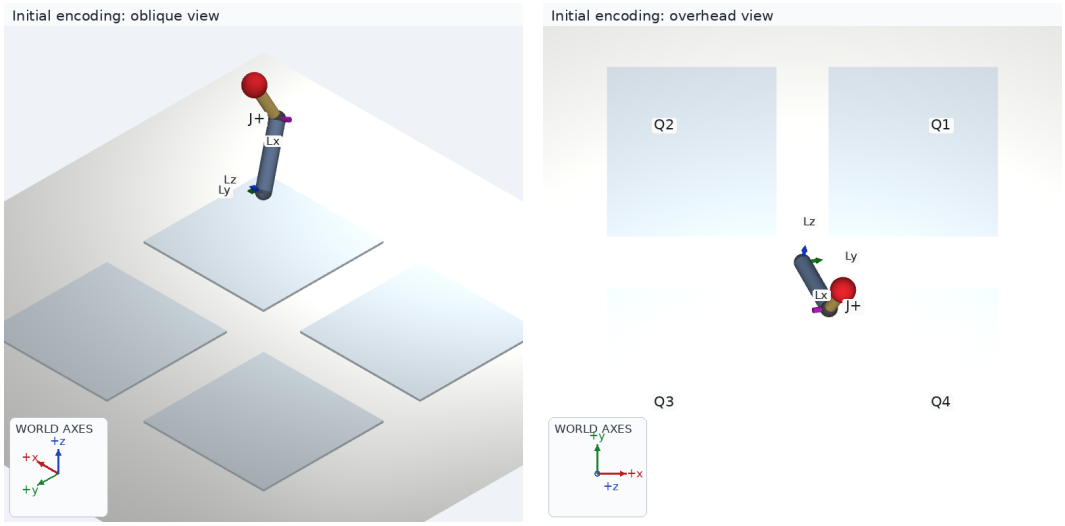
For independent stochastic draws from the same text-conditioned distribution $p(\cdot \mid x)$, the expected pair score is $p(y_0 \mid x)p(y_1 \mid x) \leq 1/4$. Uniform four-way draws give $1/16 = 6.25\%$. Thus 6.25% is a uniform-guess reference, not a universal upper bound for all text-only procedures. A pair score above this value alone does not establish visual dependence. The deterministic guarantee, language-only probes, same-input checks, and observation interventions address different aspects of

validity. These bounds assume the stated information access and draw independence; they do not cover correlated cross-question sampling strategies or label leakage.

G RENDERED DEVELOPMENT EXAMPLES

The following examples come from development roots. Their terminal states are shown only as explanatory oracle annotations; primary OL requests contain initial views and the prescribed program. No test answer is exposed by these illustrations.

G.1 FAMILY A: A FOUR-ACTION DEVELOPMENT EXAMPLE



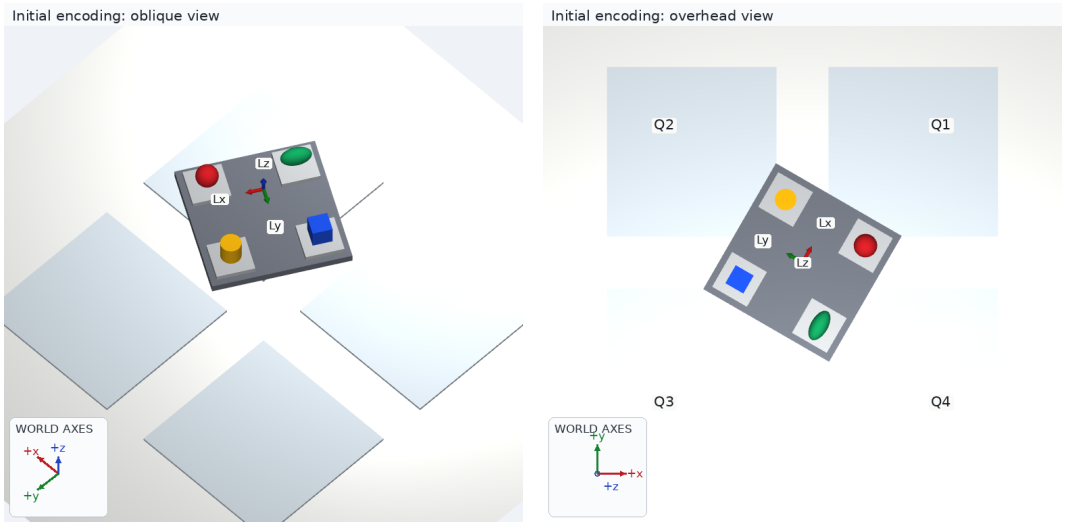
Which fixed world quadrant contains the red terminal sphere's center after all actions?

1. Increase the hinge J about root Ly angle by +30 degrees.
2. Rotate the root and all attached parts by -30 degrees about the world +z axis.
3. Increase the hinge J about root Ly angle by -60 degrees.
4. Rotate the root and all attached parts by -30 degrees about the world +y axis.

Verified development answer: Q3: world x < 0 and y < 0.


```
{"tip_world_m":[-0.058346996247578276,-0.06576345663734841,0.7430386959302258]}
```


G.2 FAMILY B: A FOUR-ACTION DEVELOPMENT EXAMPLE



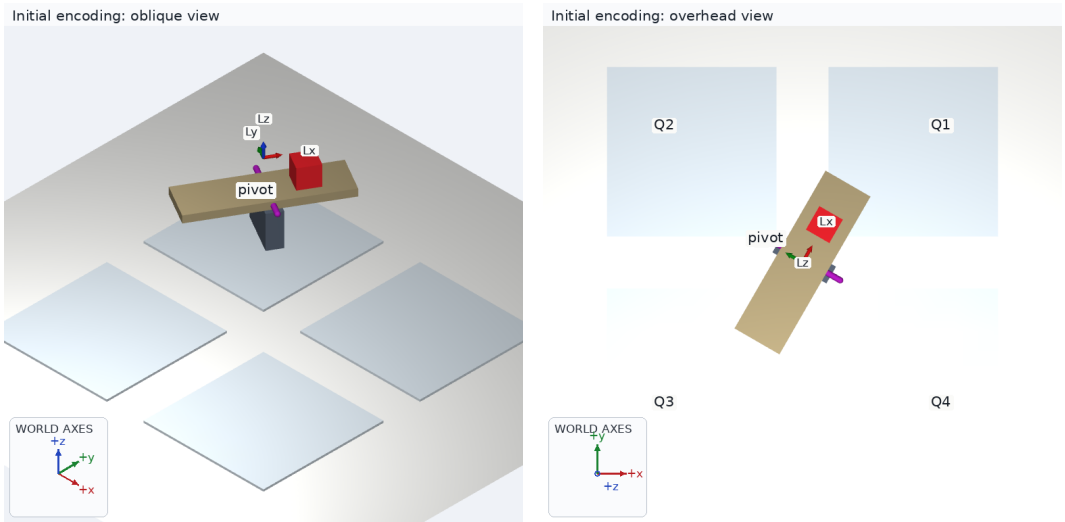
Which token's center is in fixed world quadrant Q1 after all actions? Each final quadrant contains exactly one token.

1. Rotate the root and all attached parts by +180 degrees about the local +z axis.
2. Rotate the root and all attached parts by -90 degrees about the local +y axis.
3. Rotate the root and all attached parts by +90 degrees about the world +x axis.
4. Rotate the root and all attached parts by +90 degrees about the world +z axis.

Verified development answer: yellow cylinder.


```
{"tokens_world_m":{"yellow cylinder
    ":[0.13322665703045392,0.14137766944964972,0.4784199924815638],"blue cube
    ":[-0.13322665703045386,0.14137766944964977,0.4784199924815638],"green ellipsoid
    ":[-0.13322665703045392,-0.0893776694496497,0.6116466495120177],"red sphere
    ":[0.13322665703045386,-0.08937766944964974,0.6116466495120177]}}
```


G.3 Family C: A four-action development example



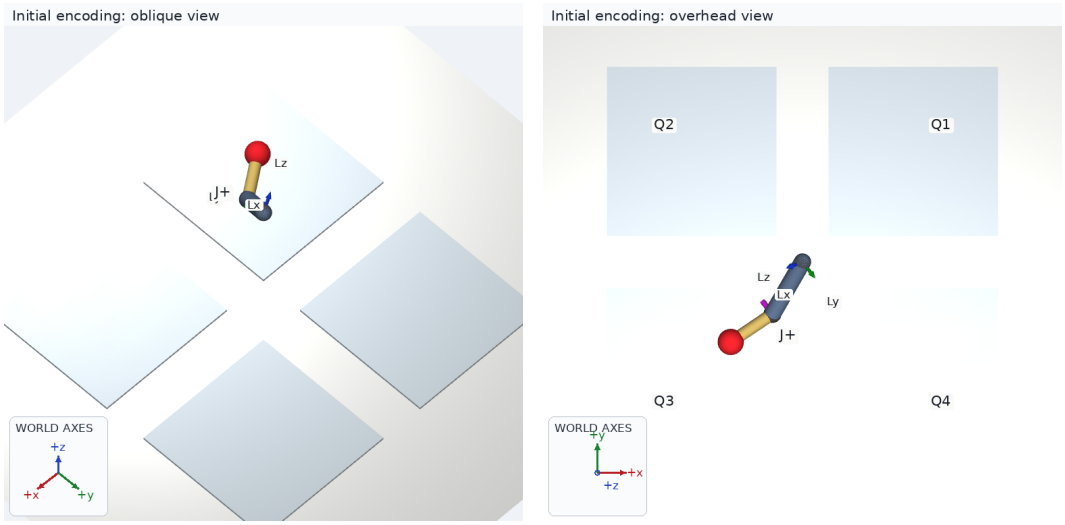
In which marked floor receiver quadrant does the red block's center come to rest after release?

1. Rotate the clamped beam assembly by -90 degrees about the world +z axis.
2. Translate the pivot by -2 cm along the beam's current Lx axis, while beam and block stay clamped.
3. Rotate the clamped beam assembly by +90 degrees about the world +z axis.
4. RELEASE the beam-level clamp and the block clamp; allow passive motion until rest (at most 5 seconds).

Verified development answer: Q1: world x > 0 and y > 0.


```
{"target_world_m":[0.15463764836770813,0.267840258043244,0.03904784952304586],"at_rest":true}
```

G.4 FAMILY D: A FOUR-ACTION DEVELOPMENT EXAMPLE



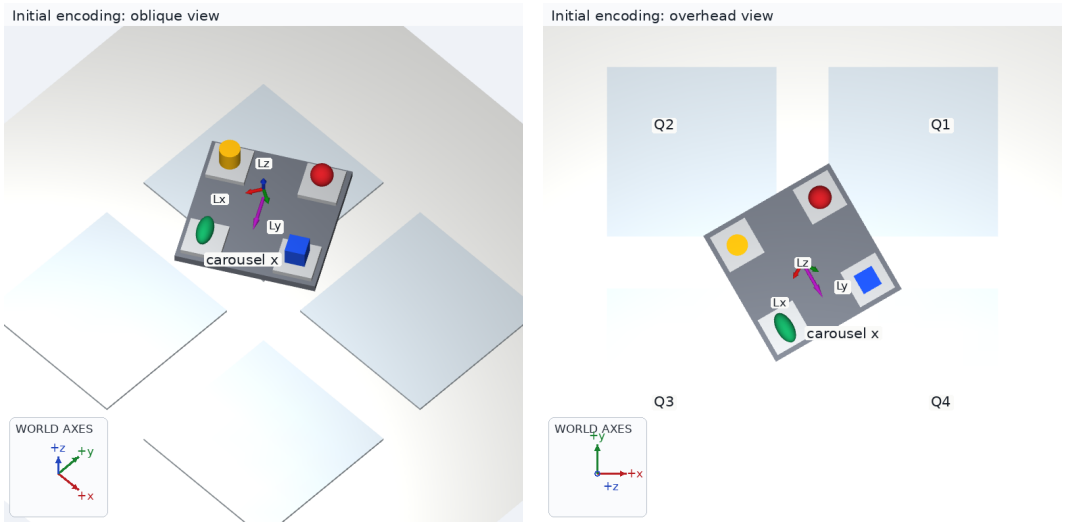
In which branch or branches does the red terminal sphere finish on the positive side of the world x=0 plane (world x > 0)?

1. Shared prefix 1: Increase the hinge J about root Ly angle by +30 degrees.
2. Shared prefix 2: Rotate the root and all attached parts by -60 degrees about the world +x axis.
3. Shared prefix 3: Rotate the root and all attached parts by -60 degrees about the world +x axis.
4. Branch A, independently from shared prefix: Rotate the root and all attached parts by -90 degrees about the local +z axis.
5. Branch B, independently from shared prefix: Rotate the root and all attached parts by -60 degrees about the world +y axis.

Verified development answer: Neither branch.


```
{"branch_A_tip_world_m":[-0.2686361879054078,-0.20302899587675788,0.32221421758418933],"
    branch_B_tip_world_m":[-0.3751977605307944,-0.06495533400528479,0.5011849578957185]}
```


G.5 FAMILY E: A FOUR-ACTION DEVELOPMENT EXAMPLE



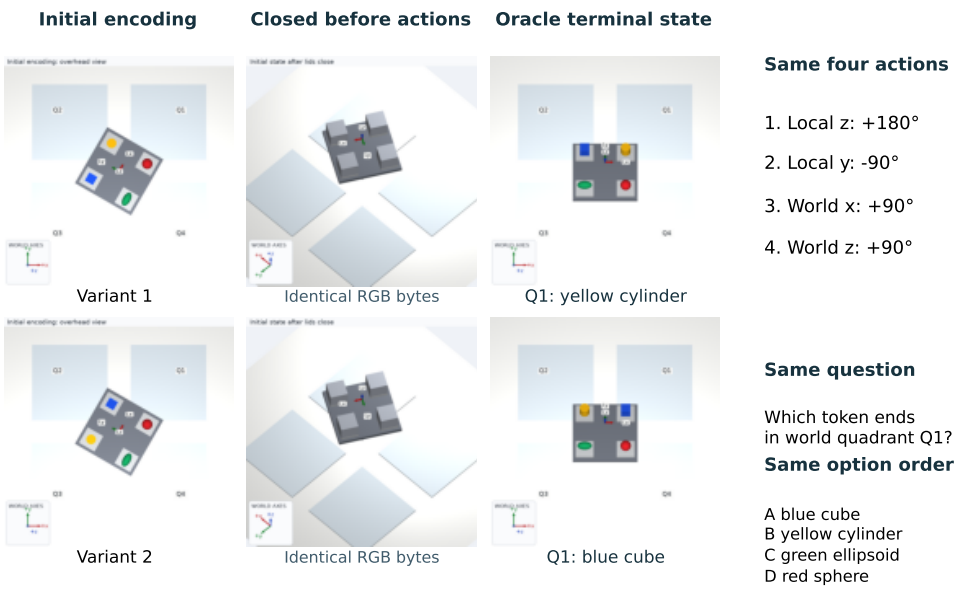


Figure 10: A worked development pair from family B. Exchanging two visible token identities changes the answer under the same program and options. The closed images are byte-identical. This compact illustration omits the additional initial oblique view and abbreviates the action wording. Revealed terminal views explain the ground truth; they are excluded from the primary input.

Which token's center is in fixed world quadrant Q3 after all actions? Each final quadrant contains exactly one token.

1. Increase the carousel joint about root Lz angle by +90 degrees.
2. Rotate the root and all attached parts by +180 degrees about the world +x axis.
3. Increase the carousel joint about root Lz angle by +90 degrees.
4. Rotate the root and all attached parts by +180 degrees about the world +y axis.

Verified development answer: green ellipsoid.


```
{"tokens_world_m":{"blue cube":[0.18843216027467236,-0.05049024517365055,0.552],"red sphere
    ":[0.05049024517365054,0.18843216027467236,0.552],"yellow cylinder
    ":[-0.18843216027467236,0.05049024517365034,0.5519999999999999],"green ellipsoid
    ":[-0.05049024517365055,-0.18843216027467236,0.5519999999999999]}}
```


H DETAILED ORIGINAL RESULTS AND FIGURE CONTROLS

Table 52 summarizes the five query families, and Table 53 identifies the information each diagnostic supplies. The full-core model comparison appears in Table 54; Figure 11 retains the separate matched-panel task and horizon profiles. Table 55 summarizes the information conditions, and Table 56 shows the effect of changing both answer instructions and output limits. Table 57 keeps independently generated shifts outside the core score. Table 58 reports stricter multi-question success requirements, so its columns should not be interpreted as interchangeable single-question accuracies. All 4,000 core labels passed post-generation re-execution, together with 1,600 branch-suffix checks and 800 physical checkpoint replays. The physical questions retain 17,600 additional stability runs. The maximum recorded penetration was 1.964 mm under the 2 mm gate. Minimum accepted target-witness area and extent were 715 pixels and 25 pixels; the minimum qualifying visible fraction was 60.79%. All 40 selected regeneration questions reproduced their complete episode records exactly. The core accepted 2,000 roots from 16,722 candidate attempts (11.96%). This is a filtered procedural distribution, not uniform sampling over 3D scenes. A separate visibility-aware finite-state

Family	Mechanism and actions	Four-way query	Answer-changing twin
A: articulation	Root rotations and a child hinge about root $+y$	Terminal marker quadrant	Initial root orientation
B: hidden identity	Covered four-token carrier; root rotations	Token in a fixed quadrant	Two token identities exchanged
C: support release	Clamped yaw/pivot preparation; passive release	Block's resting floor quadrant	Block offset reflected
D: branches	Articulated shared prefix, two independent suffixes	Positive world x: A only, B only, both, neither	Initial root orientation
E: combined updates	Covered carousel; root and child z rotations	Token in a fixed quadrant	Two token identities exchanged

Table 52: Implemented core tasks. E combines hidden identity and articulation; its current grammar does not include contact dynamics. C is one support-release mechanism, not a general physical-reasoning suite.

Condition	Information supplied	Interpretation of comparison
OL	Initial images and full program	Primary prediction task
FB	OL plus images after actions $1, \ldots, T-1$	Benefit from nonterminal visual feedback
REP	OL plus repeated initial images in the same number of slots	Control for additional images and view slots
S0	Exact initial numeric state replaces initial images	Performance with initial geometry/identity supplied
RESET	OL plus exact state after $\lfloor T/2 \rfloor$ actions	Performance after a privileged checkpoint
FINAL	Query-sufficient terminal numeric state; no images or program	Terminal readout and response-format check

Table 53: Information interventions share the same semantic questions and option order. State conditions are privileged diagnostics, not equivalent visual inputs. At $T = 1$, OL, FB, REP, and RESET have identical requests and reuse one response.

audit flagged 0 of 3,200 nonphysical core questions, using visible label anchors and target silhouettes under the fixed rendering model. Independent readout from the supplied terminal coordinates reproduces all 1,600 FINAL packet labels (1,200 distinct packets, including the pilot).

Development-trained full-text, option-layout, and action-text probes obtained 25.55%, 25.00%, and 25.65% core accuracy, respectively. Their 95% interval upper bounds were 26.50%, 25.00%, and 26.60%. The audited development/core inputs had 0 exact cross-split duplicates under the stated semantic fingerprint. The privileged static-state baseline scored 33.03 [31.40, 34.65]%, while the complete-state executor scored 100.00%. These baselines have access to exact initial state. The executor uses separate quaternion/engine kinematics and reruns the declared simulator for C.

Qwen3.8-27B-FP8 completes all 4,000 questions with 31.63 [30.28, 32.93]% accuracy, 19.40% pair accuracy, and 57.05% invalid responses. Astra scores 98.63 [98.25, 98.98]% accuracy and 97.35 [96.65, 98.00]% pair accuracy, with 0.00% invalid responses. Both runs have zero missing questions and zero infrastructure errors. The inference settings differ, so the score gap cannot be attributed to model identity alone.

Qwen3.5-9B has an FB-minus-OL difference of -2.25 [-5.50, 1.00] percentage points and an FB-minus-REP difference of -2.88 [-6.00, 0.25] points. InternVL3.5-8B has an FB-minus-OL difference of -0.25 [-2.88, 2.38] percentage points and an FB-minus-REP difference of -1.00 [-4.25, 2.25] points. These comparisons concern access to external images or exact state; they do not identify a unique internal error mechanism.

SmolVLM returned a bare option letter without the required `FINAL:` prefix on all 400 direct OL questions. The prespecified parser assigns these responses zero credit; this score records a format failure. Its direct FINAL invalid rate is 80.75%. Qwen3.5-9B has a matched FINAL direct-minus-brief difference of -38.25 [-43.50, -32.75] percentage points. InternVL3.5-8B has a matched FINAL

Model	N	Accuracy [95% CI]	Pair	Invalid
Qwen3-VL-4B	4000	25.88 [24.55, 27.18]	6.15	3.15
Qwen3-VL-8B	4000	27.15 [25.80, 28.45]	6.90	1.18
Qwen3.5-4B	4000	26.20 [24.90, 27.53]	9.80	28.43
Qwen3.5-9B	4000	33.40 [31.93, 34.90]	14.80	11.15
InternVL3.5-8B	4000	26.13 [24.95, 27.30]	4.95	0.68
SmolVLM2-2.2B	4000	22.35 [21.48, 23.23]	2.15	11.80
Qwen3.8-27B FP8	4000	31.63 [30.28, 32.93]	19.40	57.05
GPT-6 Astra (Codex, high)	4000	98.63 [98.25, 98.98]	97.35	0.00

Table 54: Scores are percentages. All completed rows use the same 4,000 OL questions and the frozen terminal-answer parser. Invalid answers count as incorrect. Astra is a separate high-reasoning Codex condition; its reasoning budget, image processing, and harness differ from the open-model condition. All answers are scored against simulator ground truth. Intervals are 95% bootstrap intervals.

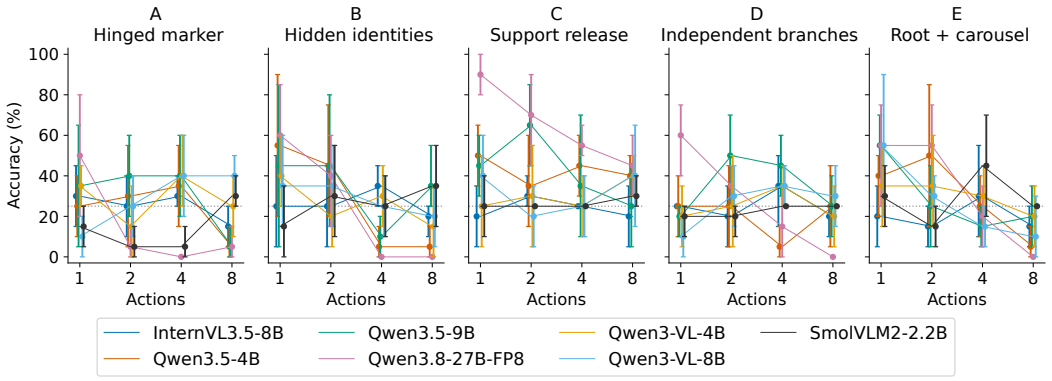


Figure 11: Original seven-open-model comparison on the matched 400 questions. Each task/horizon cell contains 20 questions. Intervals are pointwise 95% balancing-group intervals; connected horizons use independent scenes. Astra was evaluated only on the full-core OL contract and is not part of these historical intervention comparisons.

Model	OL	FB	REP	S0	RESET	FINAL
Qwen3.5-9B	33.63	31.38	34.25	29.38	32.13	95.38
InternVL3.5-8B	26.63	26.38	27.38	26.88	27.25	98.75

Table 55: Matched observation accuracy (percent), 800 questions per condition and model. Identical one-step requests reuse one response. Full intervals and paired differences, including the $T > 1$ restriction, are in the appendix.

Model	OL [95% CI]	OL invalid	FINAL	Direct − brief
Qwen3-VL-4B	25.25 [21.75, 28.75]	0.50	45.25	-2.75 [-8.00, 2.50]
Qwen3-VL-8B	28.25 [24.75, 31.75]	0.00	49.50	-0.25 [-5.75, 5.25]
Qwen3.5-4B	27.25 [24.25, 30.25]	0.25	49.00	-1.50 [-6.25, 3.50]
Qwen3.5-9B	11.00 [9.00, 13.00]	67.00	55.75	-23.00 [-28.50, -17.50]
InternVL3.5-8B	25.75 [22.50, 29.00]	0.00	42.75	1.75 [-2.50, 6.00]
SmolVLM2-2.2B	0.00 [0.00, 0.00]	100.00	4.75	-23.50 [-26.50, -20.75]
Qwen3.8-27B-FP8	2.75 [1.50, 4.25]	90.50	90.50	-30.25 [-34.00, -26.25]

Table 56: The same 400 questions under a direct-answer instruction and 32-token cap. Accuracies and OL invalid rates are percentages; the final column is the paired OL difference in percentage points from the brief-reasoning contract. Both instruction and budget change.

Model	Core $4, 8$	Geom.	Camera	Appear.	Magnitude	Stress
Qwen3.5-9B	25.20	27.90	23.00	25.90	26.30	17.00
InternVL3.5-8B	24.15	27.40	25.60	24.20	25.30	24.13

Table 57: Accuracy in percent. The core reference uses 2,000 questions at $T = 4, 8$; each separately generated shift bank has 1,000 at those horizons. Stress has 800 at $T = 12, 16$. The camera shift retains the overhead anchor. Extensions are excluded from the core aggregate.

Model	Evidence all 3	Branch mid pair	Branch all 6	Padding Δ
Qwen3.5-9B	23.50	12.18	0.86	-4.00
InternVL3.5-8B	2.00	1.75	0.00	-12.00

Table 58: Auxiliary root scores in percent; padding is padded-minus-short accuracy in percentage points. Evidence has 200 roots, branches 100, and depth controls 100. Constant abstention scores zero on all-three evidence correctness; constant branch answers fail intermediate-pair correctness. Independent uniform four-way guesses have expected branch intermediate-pair accuracy 6.25% and all-six accuracy 0.0244%.

direct-minus-brief difference of -56.25 [-60.25, -52.00] percentage points. Valid-but-wrong answers increase under the direct contract for these models. Both the instruction and output cap change.

Branch agreement is reported with correctness because a constant answer can agree after reconvergence. Hold-padding comparisons use identical scenes and terminal labels for the short and padded programs; high-sensitivity comparisons also change the active program.

I Prospective open-model follow-ups

I.1 Scope, allocation, and calibration

The follow-up protocol was fixed on 12 September 2026 after inspecting the historical benchmark results, and is therefore exploratory. It specifies Qwen3.5-9B and InternVL3.5-8B, using their previously pinned checkpoints and serving configuration. The common panel contains the same 400 questions used for the seven-model comparison. All variants and original dependency groups are retained. Each model is allocated 7,004 diagnostic request records: 4,800 strategy, coordinate, calibration, and tool calls; 1,584 prefix calls; 400 checkpoint/control calls; and 220 frame-program calls. An additional 400 ordinary OL requests use the unchanged evaluator at the follow-up serving concurrency and seed, giving 7,404 scored request records per model. This fresh baseline was allocated before follow-up accuracy analysis; the historical seven-model comparison is retained separately. Four-sample voting retains every individual response. The full run is not considered complete while any allocated request is absent or has an unresolved infrastructure error.

Absolute coordinate errors require a declared metric reference. For coordinate and tool prompts, the rule card therefore adds the hub position and static mechanism dimensions: arm-link lengths for A/D; carrier spacing and token-center height for B/E; and beam/block dimensions for C. Initial orientation, joint angle, pivot displacement, and token assignment remain hidden. A geometry-only final-answer condition receives the same calibration. This control separates the additional information from the rest of the tool intervention. An initial prototype lacking this calibration was stopped before scoring. Its responses and source version are retained separately and excluded from all follow-up figures. The correction followed a methods review, not item accuracy.

A second pre-analysis review found that the coordinate instruction listed token names in hidden compartment order. Version 3 sorts the requested names lexicographically, independently of the scene. An automated check verifies identical name ordering and calibration for every pair of variants. Affected prototype requests are excluded and rerun. An earlier response can be reused only when its complete request hash, saved body, image references, pinned model/serving settings, and dependency metadata match the corrected request. Reused raw bytes are unchanged and have indi-

vidual source receipts. This avoids reexecuting unaffected controls while preserving the correction's provenance.

I.2 Coordinate prediction and prefix survival

A deterministic hash ordering selects two complete balancing groups in each of the 20 task/horizon cells: 160 questions in total. The 792 query states include initial states, each action boundary, and both independent D endpoints. C's final query uses the stored settled block position. D branches are evaluated from the same shared-prefix state and never chained together.

The requested JSON contains the root rotation matrix R, hub position, joint angle in degrees, pivot offset in meters, and named world-coordinate points. The coordinate parser selects the first decodable JSON object containing a `points` dictionary, including objects surrounded by prose or code fences. It does not repair numeric values or select among candidates by correctness. Arms require elbow and tip; carriers require the four named token centers; C requires the block center. Terminal D queries instead require both branch tips. Each independent prefix query starts without earlier model outputs. The sequential condition additionally receives only the preceding predicted JSON, or an explicit invalid-state marker. It retains the full prescribed prefix, so the model can recompute it; this is not a forced incremental simulator. The two initial queries have identical information and are executed separately. Prefix requests retain the original final-question options as a common scaffold; those options need not describe the intermediate state.

For structurally valid states, position error is the mean Euclidean error across required named points. Position success requires every required point to be within 2 cm. Identity swaps therefore fail even when the unordered point set is correct. Root-rotation error is the geodesic angle between rotation matrices; joint error uses the shortest periodic angular difference. Their success thresholds are five degrees. Structural validation requires finite three-vectors with coordinates within $[-3, 3]$ meters, a finite 3×3 matrix with $\|R^\top R - I\|_F \leq 0.15$ and $|\det R - 1| \leq 0.1$, and the required finite scalars. Invalid states count as failures; continuous errors are conditional on valid output. The settled C endpoint and terminal D joint query are assessed through their query-relevant points rather than an invented dynamic root pose. These are declared numerical diagnostics, not human-validated perceptual tolerances.

The headline prefix plot fixes nominal horizon at eight actions, retaining 40 questions across all five tasks. D's two final endpoints are averaged within item before root aggregation. Correct-prefix survival requires success at every queried predecessor and the current state; each D path includes only its own endpoint. Independent-query survival measures consistency across separate answers, not error accumulation in an autoregressive state trajectory. All other horizon/task cells and validity rates are retained in the score artifact. A companion curve reports mean coordinate error in centimeters among valid states. Its contributing roots and task cells can change with output validity; it is interpreted together with the unconditional success and validity curves.

I.3 Checkpoint repair

For each selected item, checkpoints are placed at $\lfloor T/2 \rfloor$ and $T - 1$, with duplicates removed and $T = 1$ excluded. The exact-state condition supplies the verified nonterminal state and instructs the model to advance the suffix. Its control identifies the same checkpoint slot without supplying state values. The headline contrast includes every assigned item, regardless of earlier correctness. At $T = 2$, the common position contributes to the midpoint category only. The two position categories consequently have different memberships; comparing their magnitudes does not isolate a pure checkpoint-position effect.

I.4 Equivalent frames and action order

A separate search uses only geometry and labels to construct two-action programs whose order reversal changes the answer while both endpoints retain unique quadrants and the 2 cm margin. It searches fixed angle and axis candidates in a deterministic hash order, independently of model responses. Both paired variants must satisfy the criterion, and a source balancing group is retained only when both of its roots pass. The accepted bank has 44 questions in 11 groups: 36 B questions and eight E questions. This is smaller than the target of 20 groups; all accepted groups are retained,

with no accuracy-based expansion or removal. The B groups come from source horizons one, two, four, and eight (two, two, three, and two groups); E contributes two groups, both from source horizon eight. These are source strata: every new noncommuting program itself has two actions.

For a world-axis rotation with direction u, the equivalent current-local axis is $R^\top u$. The local action is then composed on the right, and the next axis is computed after updating R. The public wording supplies the three numeric components of this unit vector. An independent parser re-executes that wording with SciPy rotations; MuJoCo landmarks check the endpoints. These coefficients can reveal directional information about the initial pose, so the comparison also changes representation and available numeric information. The same coefficients apply to both identity variants of a root, so they do not disclose which token occupies a compartment.

A commuting control splits the first rotation into one-third and two-thirds of its angle, then swaps those two parts while retaining the second original rotation. This yields two equivalent three-action programs. We compare those programs within their shared length; the noncommuting pair contains two actions. We report both agreement and joint correctness. Agreement requires a valid answer in both requests. All 11 accepted groups retain each display letter and semantic answer once in every program condition; no further rebalancing was needed. Condition-specific label histograms accompany the audit. Source dependency groups are retained for uncertainty. The small and uneven task coverage limits inference about general reference-frame competence.

I.5 Prompting, sampling, and numeric tools

The unchanged brief instruction is evaluated at caps 512 and 4,096, alongside the fresh 2,048-token baseline. The state-table condition requests a concise per-action table within the same brief contract. Demonstrations are two fixed textual development examples for the item's task, at horizons one and two, including their initial numeric description, program, options, and verified terminal state. Demonstration membership is selected by a fixed hash rule and never by test performance.

Four stochastic calls use temperature 0.7 and distinct fixed seeds. The voting rule selects the most frequent valid option; ties follow the earliest sampled valid option among the tied choices. Four invalid samples yield an invalid aggregate. Report the four-sample result against both the greedy baseline and the first stochastic sample. Any-of-four correctness uses an oracle selector and is explicitly excluded from deployable-strategy accuracy. Request records retain the exact transmitted settings and the checkpoint's inherited defaults.

The tool condition first asks the model to estimate initial coordinates and pose from the calibrated initial views. A fixed numeric executor then transforms those predictions using the declared actions. Arm joints rotate the predicted tip about the predicted elbow and root-y direction; carousel joints rotate predicted token centers about the predicted root-z direction. D branches start independently from the predicted prefix. C uses a fresh simulation initialized from predicted block/pivot offsets, height, and yaw; only the beam/block dimensions are taken from the disclosed rule card. A near-yaw matrix is projected to its yaw constraint. Invalid inputs, unsupported poses, or failed physical rollouts return an explicit tool error rather than oracle values. The physical initializer places the block on the beam centerline at its estimated local-x offset, using the estimated hub height and declared object dimensions; off-axis block coordinates are not simulated as free-space offsets. For C, the returned JSON updates the block's settled world position and retains the other fields as estimated input metadata, not a settled beam-pose estimate.

A second model call receives the executor result or error and answers the original question. Thus the tool intervention has two model calls and additional calibration; its paired geometry-only control is essential. Exact-input tool tests reproduce all 400 frozen answers, with maximum point discrepancy below 10^{-12} m. This checks implementation consistency, not the accuracy of model-extracted state. Token totals include both model calls. Model-call latency is recorded, but tool computation time was not separately measured, so these records do not establish end-to-end speed comparisons.

I.6 State–answer coupling and reproducibility

A distinct calibrated terminal query requests a JSON state followed by the final option. The frozen geometric readout is applied to the predicted points, using a small numerical boundary guard and requiring unique token quadrants. The result and the model's own option define four categories: both

correct, state only, answer only, and neither. Missing or unusable coordinates fail the state-derived answer. No generated images or hidden reasoning states are measured by this comparison.

Analyses preserve source balancing groups, use 10,000 task/horizon-stratified bootstrap draws, and report pointwise intervals. New frame programs retain the source-horizon strata for resampling; sparse strata and unequal task counts are disclosed. Repeated samples and prefix states are grouped with their original items rather than treated as independent observations. Available state errors are averaged within each root before task/horizon macro-averaging. Continuous summaries therefore describe roots with valid outputs, rather than pooling every predicted point; invalid predictions remain failures in the success-rate metrics. The intervals condition on the recorded samples and serving runs. No family-wise significance claims are made.

The source registry records summary hashes and matched membership. Each figure exports its plotted CSV, caption, vector PDF/SVG, raster preview, and source-hash manifest. The static HTML figure gallery requires no JavaScript or HTTP server. Raw requests use content-addressed image references; full model responses, request hashes, seeds, failures, and tool results remain in the local archive. Additional checkpoints can be registered after checking exact sample membership and the common response contract, then plotted without rerunning older models.

I.7 Post-hoc structured-output diagnostic

The original coordinate contract can fail before a complete state is emitted. After inspecting Qwen3.5-9B's v3 results, we allocated a separate diagnostic for both models before inspecting InternVL3.5-8B's v3 accuracy. The allocation adds 3,184 calls per model: 400 calibrated answer-only queries, 400 joint terminal-state and answer queries, 400 initial-state estimates followed by 400 tool-answer queries, and the same 792 prefix states under each of the two protocols. Five one-action development examples per model check API and JSON completeness only; they are excluded from all reported accuracy denominators.

This condition changes both instruction and decoding. The prompt requests an immediate JSON object, and the server enforces a JSON schema specifying object fields, array dimensions, numeric types, and an answer-letter enumeration. Point names remain lexicographically ordered. The schema contains no true coordinates, rotations, joint angles, identity assignments, or answer values. The calibrated answer-only control uses the same answer schema and geometry disclosure. All calls use temperature zero, a 2,048-token cap, seed 20260912, concurrency 16, and disabled native thinking where supported. JSON answers are scored by a separate parser; the original terminal-tag parser and v3 scores are unchanged.

Schema completeness does not imply a proper rotation or accurate coordinates. We retain the original geometric validity checks, 2 cm all-point success threshold, coordinate readout, and source balancing groups. Tool execution uses the model's initial estimate and the same fixed executor as v3; errors are returned to the second model call and remain in the denominator. The separate protocol, raw responses, requests, and scorer hashes are retained. This post-hoc comparison is a diagnostic of the response contract, not a preregistered test of a particular reasoning mechanism.

I.8 Post-hoc voting completeness control

Answer aggregation follows the general motivation of self-consistency (Wang et al., 2023), with the specific four-sample plurality and tie rule declared above. Using the four already recorded stochastic samples in their original order, we compare plurality selection with the first valid answer among those samples. All four calls were executed, so this analysis does not establish the cost of an early-stopping policy. We also divide the total vote-minus-first-sample gain into contributions from questions whose first response was valid or invalid. Both contributions use all 400 questions as their denominator and sum to the total gain. These validity strata are descriptive, not randomized groups. The primary sampling results remain unchanged.

I.9 Reading the intervention results

Tables 59 and 60 report absolute accuracy and paired effects for the same question panel. The fresh brief baseline controls the serving configuration used for the follow-ups; the earlier baseline remains

visible to show run dependence. Figure 27 presents the absolute scores, and Figure 6 compares each intervention with its relevant control. Figure 15 shows that response validity and length termination can change with these interventions.

Table 61 and Figure 26 distinguish deployable plurality voting from oracle selection among four samples. Table 74 and Figure 29 then compare voting with simply choosing the first valid answer. For Qwen3.5-9B, that completeness control accounts for much of the gain over the first sample; the remaining plurality benefit is smaller. Table 68 and Figure 28 count every sampled or tool-related model call. Their token totals describe these runs, not hardware-independent computational cost.

Geometry, output validity, and checkpoints. Table 75 explains why the original state queries cannot cleanly isolate geometric errors: most Qwen3.5-9B terminal outputs stop at the cap. Figures 23 and 19 separate coordinate success from usable poses. Figure 22 conditions continuous distance on valid output, whereas Figure 20 requires every preceding state to succeed. Changing valid-output membership prevents reading the distance curve as a trajectory for one fixed set of successful predictions.

Table 62 and Figure 21 compare exact checkpoints with matched scaffolding. The last-nonterminal contrast is positive for both models, but its membership differs from the midpoint contrast. This rules out attributing their difference solely to checkpoint timing. Under guided JSON, Figure 5 still shows low initial-coordinate success; Figure 24 reports geometric validity and conditional distance separately.

State answers and fixed execution. Tables 65 and 66, together with Figure 31, compare coordinate-derived options with separately emitted answers. Agreement can be wrong, and incomplete state outputs remain failures. Table 67 shows that the fixed executor often cannot execute the extracted state; its failures remain in the accuracy denominator and reach the final model call as errors.

The guided results keep a separate response contract. Table 69 compares answer-only, state-and-answer, and tool prompts; Table 70 confirms complete JSON for these requests. Table 71 shows that complete JSON still permits invalid geometry and execution failures. The tool-minus-answer-only intervals in Table 72 include zero for both models. Table 73 and Figure 32 report joint coordinate-readout and answer outcomes. Here state correctness means the coarse answer obtained from coordinates, not an accurate full pose.

Task structure and generalization. Figure 12 contrasts single-question and paired correctness on the matched open-model panel. Figure 13 and Figure 14 expand the information comparisons by task; the latter shows why an aggregate feedback score can hide variation across families. Figure 16 requires correctness on related scenes or branches, while Figure 17 separates answer prediction from detecting insufficient evidence. Figure 18 places hold-padding performance beside measured single-action sensitivity: program length is not a certificate of minimum reasoning depth.

Tables 63 and 64, and Figure 25, distinguish equivalent-program agreement from joint correctness. The small frame bank covers only carrier families; numeric axis coefficients also change the information supplied. The result therefore has narrower scope than a general test of reference-frame invariance. Finally, Figure 30 compares the shift banks with a core reference at matching horizons and reports stress separately. Its near-chance scores describe these two open models, not Astra, which was evaluated only on the core OL condition.

For Qwen3.5-9B, the returned terminal-state points have a usable-state rate of 0.25 [0.00, 0.75]% and all-point coordinate success of 0.25 [0.00, 0.75]%. Continuous error summaries condition on valid state outputs and are available with all task/horizon cells in the machine-readable score artifact.

For InternVL3.5-8B, the returned terminal-state points have a usable-state rate of 54.25 [49.75, 58.75]% and all-point coordinate success of 0.00 [0.00, 0.00]%. Continuous error summaries condition on valid state outputs and are available with all task/horizon cells in the machine-readable score artifact.

Condition	Qwen3.5-9B	InternVL3.5-8B
Historical 2048	34.00 [29.00, 38.75]	24.00 [21.00, 27.25]
Fresh 2048	30.25 [25.75, 34.75]	25.75 [22.25, 29.25]
Cap 512	26.75 [22.75, 30.75]	24.75 [21.25, 28.25]
Cap 4096	34.50 [29.75, 39.25]	22.25 [19.00, 25.50]
State table	34.75 [30.50, 39.00]	27.00 [23.25, 30.75]
2 demos	34.75 [30.75, 38.75]	28.50 [25.25, 32.00]
1 sample	30.00 [26.25, 33.75]	27.00 [23.25, 30.75]
4-sample vote	40.25 [35.50, 45.00]	25.75 [22.50, 29.00]
Numeric tool	39.50 [35.25, 43.75]	31.00 [27.50, 34.50]
Geometry only	34.00 [29.50, 38.50]	27.75 [24.75, 30.75]

Table 59: Follow-up accuracy on the matched 400, in percent with pointwise 95% intervals. Geometry-only calibration is an information control; tool calls use model-estimated initial state.

Condition	Qwen3.5-9B	InternVL3.5-8B
Historical 2048	3.75 [-0.25, 8.00]	-1.75 [-4.75, 1.25]
Cap 512	-3.50 [-7.25, 0.25]	-1.00 [-3.50, 1.50]
Cap 4096	4.25 [0.50, 8.00]	-3.50 [-6.00, -1.25]
State table	4.50 [-0.50, 9.50]	1.25 [-3.00, 5.75]
2 demos	4.50 [-1.00, 9.75]	2.75 [-1.50, 7.00]
1 sample	-0.25 [-5.75, 5.25]	1.25 [-3.50, 6.00]
4-sample vote	10.00 [4.50, 15.75]	0.00 [-4.00, 4.25]
Numeric tool	9.25 [4.00, 14.75]	5.25 [0.00, 10.50]
Geometry only	3.75 [-2.00, 9.50]	2.00 [-1.50, 5.50]
Tool minus geometry	5.50 [0.50, 10.50]	3.25 [-1.25, 8.00]
Vote minus 1 sample	10.25 [6.75, 14.00]	-1.25 [-3.75, 1.26]

Table 60: Paired differences in percentage points. All rows except Tool minus geometry and Vote minus 1 sample subtract the fresh brief 2,048-token baseline. All contrasts are exploratory and pointwise.

Condition	Qwen3.5-9B	InternVL3.5-8B
Any of 4 (oracle)	66.75 [62.25, 71.00]	52.75 [48.50, 57.00]
4-sample plurality	40.25 [35.50, 45.00]	25.75 [22.50, 29.00]
Sample 1	30.00 [26.25, 33.75]	27.00 [23.25, 30.75]
Sample 2	31.25 [27.00, 35.50]	22.50 [19.50, 25.50]
Sample 3	27.25 [23.00, 31.50]	27.00 [23.50, 30.25]
Sample 4	28.00 [24.25, 32.00]	23.25 [19.50, 27.50]

Table 61: Stochastic samples and valid-answer plurality on the same questions. Any-of-four uses an oracle selector and is not deployable model accuracy. Rates and intervals are percentages.

Condition	Qwen3.5-9B	InternVL3.5-8B
midpoint	1.67 [-5.83, 9.17]	4.17 [-0.83, 9.17]
last nonterminal	12.50 [1.25, 23.75]	8.75 [1.25, 16.25]

Table 62: Exact checkpoint minus matched checkpoint scaffolding, in percentage points. The midpoint and last-nonterminal categories have different memberships because coincident positions are deduplicated.

Condition	Qwen3.5-9B	InternVL3.5-8B
original	17.50 [10.00, 25.00]	10.83 [2.50, 20.00]
equivalent	23.33 [11.67, 35.00]	22.50 [12.50, 32.50]
swapped	29.17 [25.00, 33.33]	9.17 [2.50, 15.83]
commuting	22.50 [15.00, 30.00]	26.67 [21.67, 32.50]
commuting_swapped	19.17 [12.50, 25.83]	30.83 [15.83, 45.83]

Table 63: Frame-bank condition accuracy, in percent. The bank contains 44 questions and has unequal task frequencies but retains four-way label balance within every group; source dependency groups are preserved.

Condition	Qwen3.5-9B	InternVL3.5-8B
Equivalent agree	14.17 [5.00, 23.33]	34.17 [24.17, 44.17]
Equivalent both correct	2.50 [0.00, 5.00]	1.67 [0.00, 5.00]
Changed order both correct	8.33 [3.33, 13.33]	0.00 [0.00, 0.00]
Commuting agree	31.67 [25.00, 39.17]	55.00 [47.50, 62.50]
Commuting both correct	5.83 [1.67, 10.00]	17.50 [8.33, 26.67]

Table 64: Consistency and joint correctness, in percent. Equivalent-program coefficients provide numeric direction information; commuting controls compare equal-length three-action programs.

Condition	Qwen3.5-9B	InternVL3.5-8B
Both correct	0.25 [0.00, 0.75]	7.00 [5.00, 9.00]
State only	0.00 [0.00, 0.00]	5.50 [3.75, 7.50]
Answer only	0.00 [0.00, 0.00]	17.25 [14.50, 20.00]
Neither	99.75 [99.25, 100.00]	70.25 [66.50, 74.00]

Table 65: Joint outcomes for terminal-coordinate readout and the separately emitted answer, in percent. The same 400 calibrated queries are used for both outputs.

Condition	Qwen3.5-9B	InternVL3.5-8B
Consistent	0.25 [0.00, 0.75]	28.00 [24.00, 32.00]

Table 66: Consistency between a usable state-derived option and the model answer, in percent. Consistency can be wrong.

Condition	Qwen3.5-9B	InternVL3.5-8B
executor_answer_correct	6.50 [4.50, 8.50]	16.00 [13.25, 18.75]
tool_failure	89.25 [86.50, 91.75]	40.00 [36.50, 43.75]

Table 67: The fixed executor result before the second model call: correctness and execution failure rates, in percent. Failures remain in the denominator.

Model	Condition	Model calls	Output tokens
Qwen3.5-9B	baseline	400	233417
Qwen3.5-9B	cap512	400	131745
Qwen3.5-9B	cap4096	400	292009
Qwen3.5-9B	table	400	291828
Qwen3.5-9B	demonstrations	400	216602
Qwen3.5-9B	vote0	400	315324
Qwen3.5-9B	vote4	1600	1274752
Qwen3.5-9B	tool	800	872852
Qwen3.5-9B	calibration	400	242796
InternVL3.5-8B	baseline	400	83007
InternVL3.5-8B	cap512	400	82008
InternVL3.5-8B	cap4096	400	81953
InternVL3.5-8B	table	400	112497
InternVL3.5-8B	demonstrations	400	95030
InternVL3.5-8B	vote0	400	82357
InternVL3.5-8B	vote4	1600	330642
InternVL3.5-8B	tool	800	175543
InternVL3.5-8B	calibration	400	83239

Table 68: Recorded model-call and output-token totals. Voting includes all four samples; tools include extraction and final-answer calls. Tool computation was not separately timed. No API-price or end-to-end speed comparison is made.

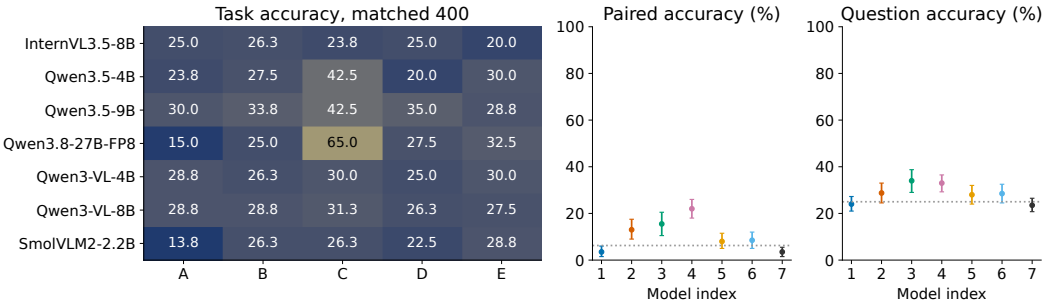


Figure 12: Question accuracy and paired correctness summarize different requirements. Heatmap and intervals use the same matched 400 questions for every model. Pair chance is 6.25% for independent uniform guesses, not a universal text-only ceiling. Model indices follow heatmap order; checkpoint sizes and serving precision differ across families.

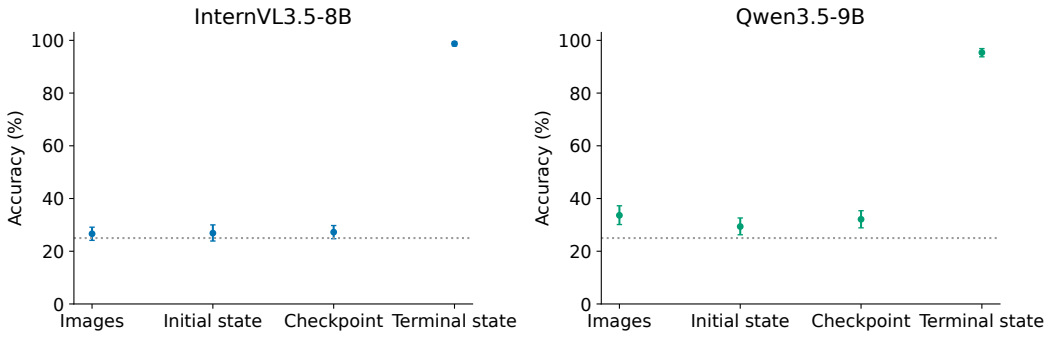


Figure 13: Terminal-coordinate readout is strong while initial-state and checkpoint assistance remain limited in these two models. All conditions use the same 800 questions per model. Exact state changes representation and precision; these contrasts do not uniquely identify internal error mechanisms. Bars denote pointwise 95% balancing-group intervals.

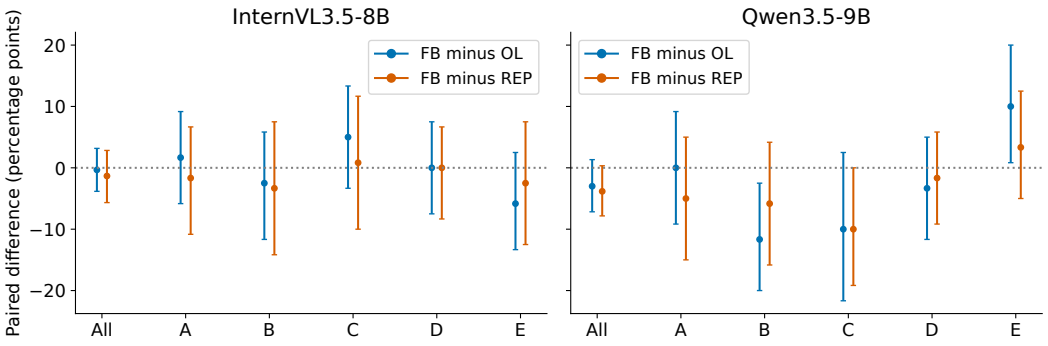


Figure 14: Nonterminal feedback does not show a reliably positive aggregate gain under the tested contract. Paired FB-minus-OL and FB-minus-REP contrasts exclude reused one-action requests. Task-specific intervals expose uncertainty and heterogeneity; FB includes the declared covered-token treatment. All intervals are pointwise 95%.

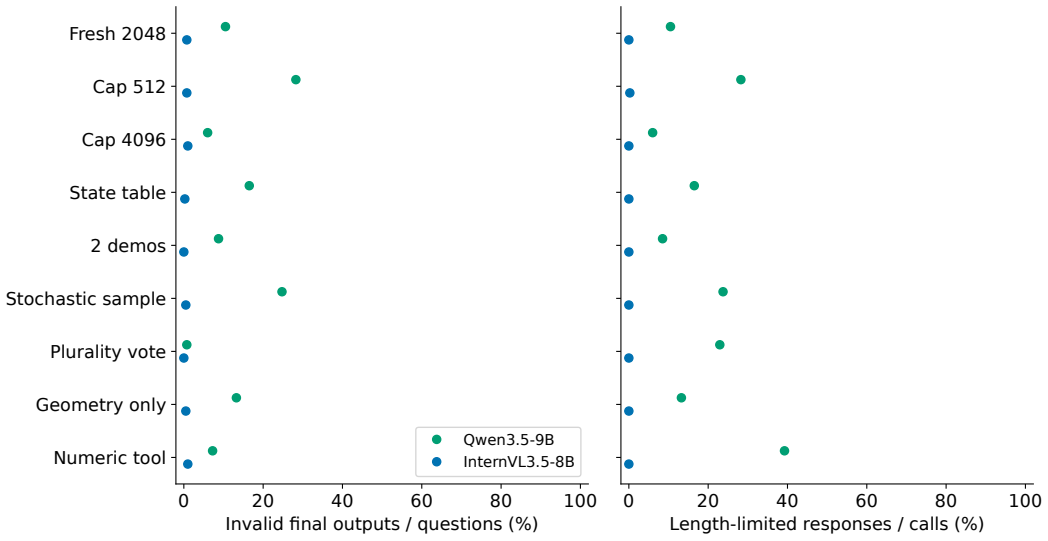


Figure 15: Response completeness accompanies the fixed-prompt cap and strategy comparison. Left: invalid final options among 400 questions, after aggregation for voting. Right: provider-reported length termination among model calls, including all four voting samples and both tool calls. These rates have different denominators and are descriptive; a truncated response need not be parser-invalid. The original answer parser is unchanged.

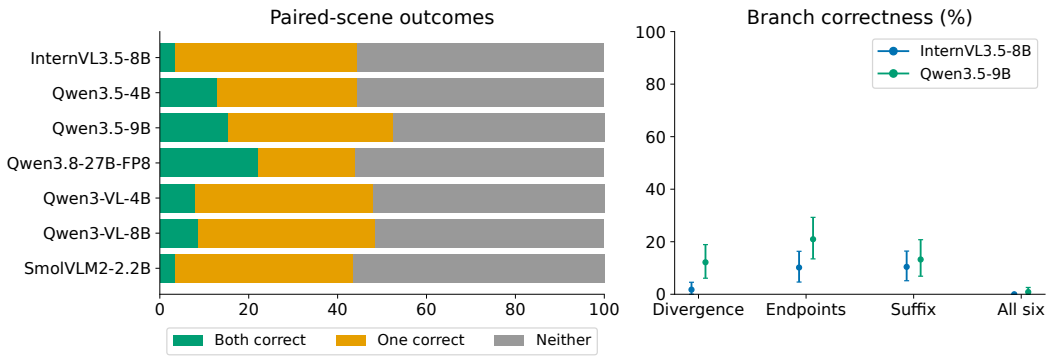


Figure 16: Solving one scene does not ensure correct tracking of its paired variant or both counterfactual branches. Left: paired outcomes on the matched 400. Right: independently queried branch diagnostics, with paired correctness at each stage and all-six correctness. Agreement without correctness can reward constant answers; branch-bank intervals retain its actual dependency groups.

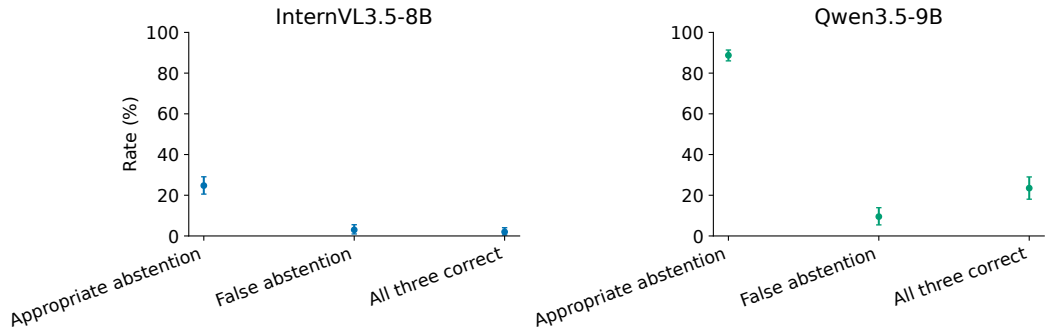


Figure 17: Evidence sufficiency must be evaluated alongside answer correctness. Appropriate abstention concerns underdetermined inputs; false abstention concerns decisive evidence. All-three correctness requires the appropriate response to every form of a scene. These diagnostics use an additional insufficiency option, so a four-option chance line would be inappropriate.

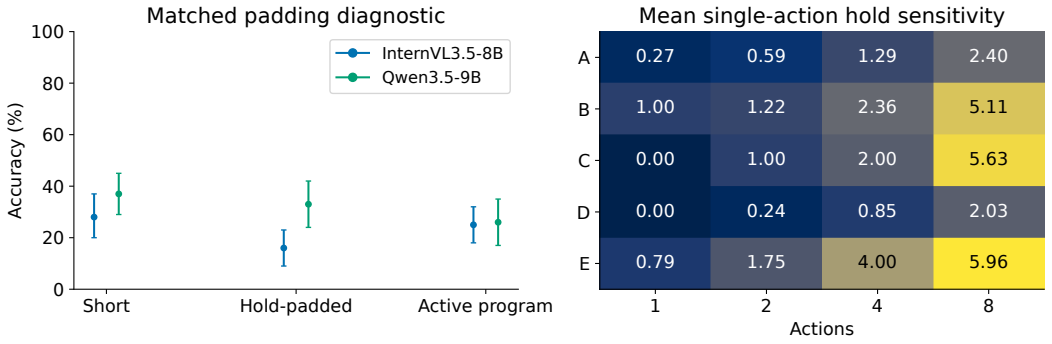


Figure 18: Program length and action dependence are distinct properties. Short and hold-padded programs preserve scene and endpoint; active programs change the computation. The sensitivity map counts individual action replacements that change the answer, not minimum reasoning depth. C release and D suffix conventions follow the dataset audit. Pointwise intervals apply to model scores.

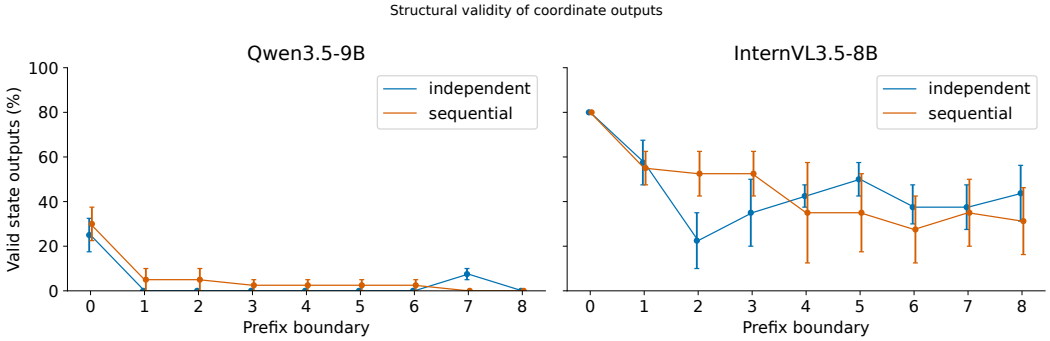


Figure 19: Structural validity is shown separately from geometric accuracy. These use the original state-output contract on the fixed $T = 8$ cohort; finite coordinates and a valid approximate pose are required. Invalid outputs remain failures in the coordinate-success plot.

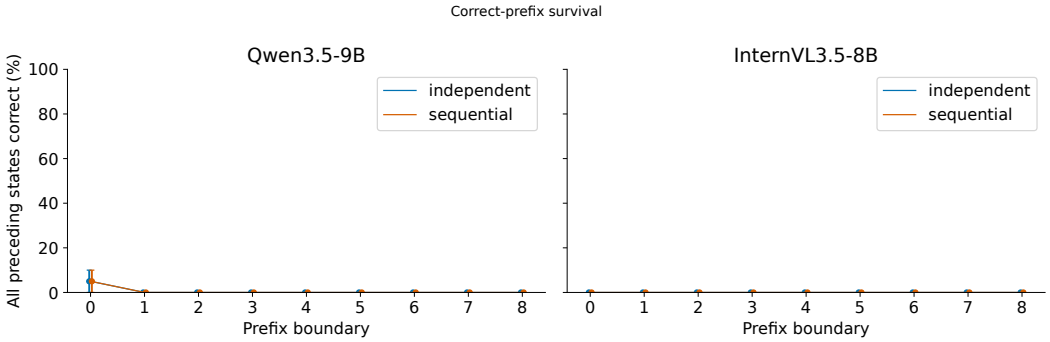


Figure 20: Correct-prefix survival requires every preceding state to pass the declared coordinate threshold. Both protocols retain original dependency groups; independent-query survival is not an autoregressive error trajectory.

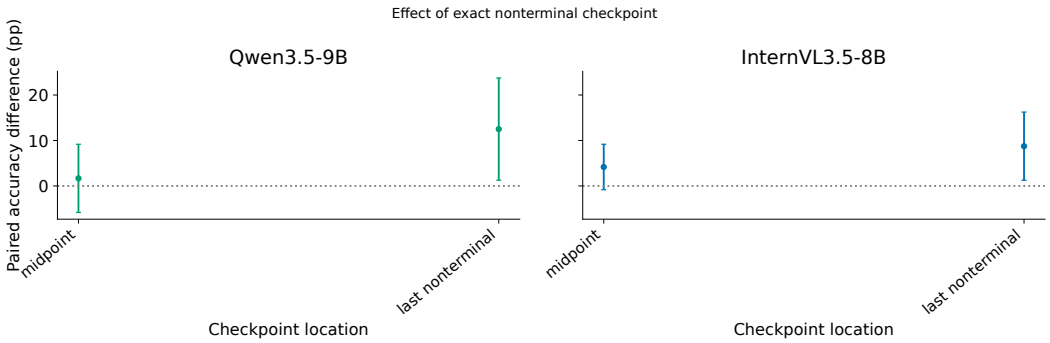


Figure 21: Verified checkpoints are compared with matched checkpoint scaffolding on every pre-selected item, without conditioning on earlier errors. Values are paired accuracy differences with pointwise 95% balancing-group intervals. Midpoint and last-nonterminal positions coincide at T=2 and are counted once; their item memberships differ.

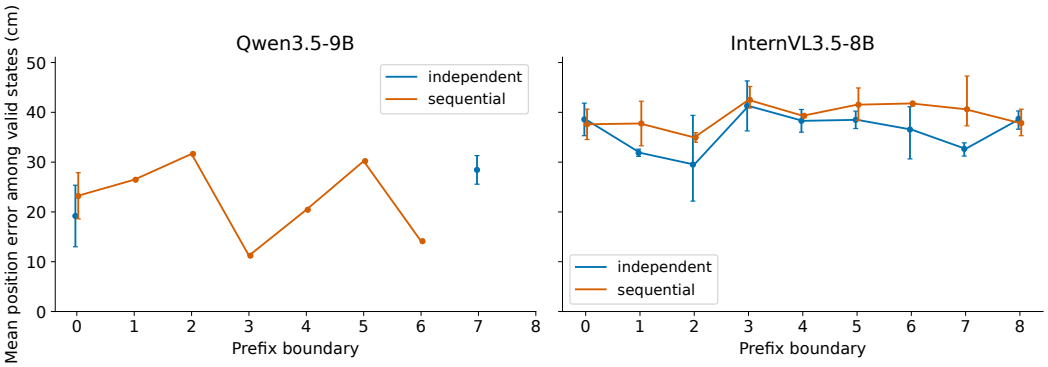


Figure 22: Continuous position errors complement the strict success criterion. The same $T = 8$ cohort is queried at every boundary. Errors average named-point distances within roots and macro-average available task cells; only structurally valid outputs contribute. The contributing questions and cells can change with validity, so this curve must be read with the success and validity panels. Missing valid outputs leave a gap rather than an invented error. Intervals preserve source dependency groups; CSV errors are in meters and the plot shows centimeters.

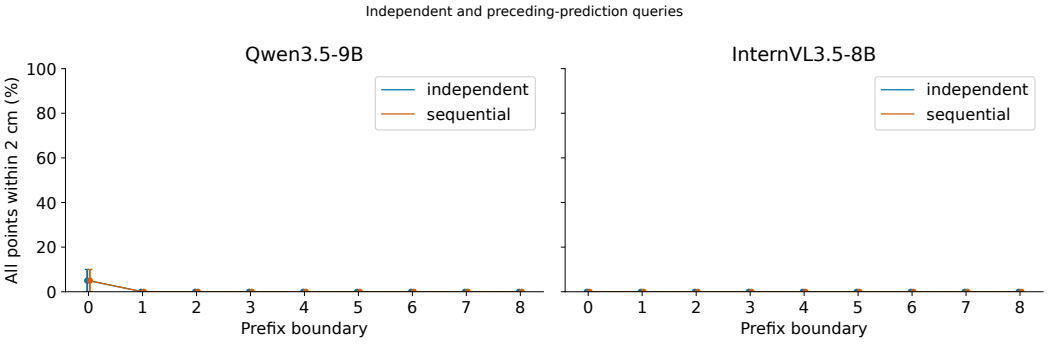


Figure 23: Original-contract independent endpoint coordinate success is Qwen3.5-9B 0.00%; InternVL3.5-8B 0.00%. Output completeness limits its interpretation. Stepwise coordinate estimates are evaluated separately from final multiple-choice answers. Independent queries start fresh; sequential queries receive only their preceding predicted state plus the full prefix. This panel fixes T=8 (40 questions), averages D branch endpoints within item, and counts invalid outputs as failures. Static dimensions and hub position are disclosed. All required named points must be within 2 cm. Pointwise balancing-group intervals do not establish an internal simulation mechanism.

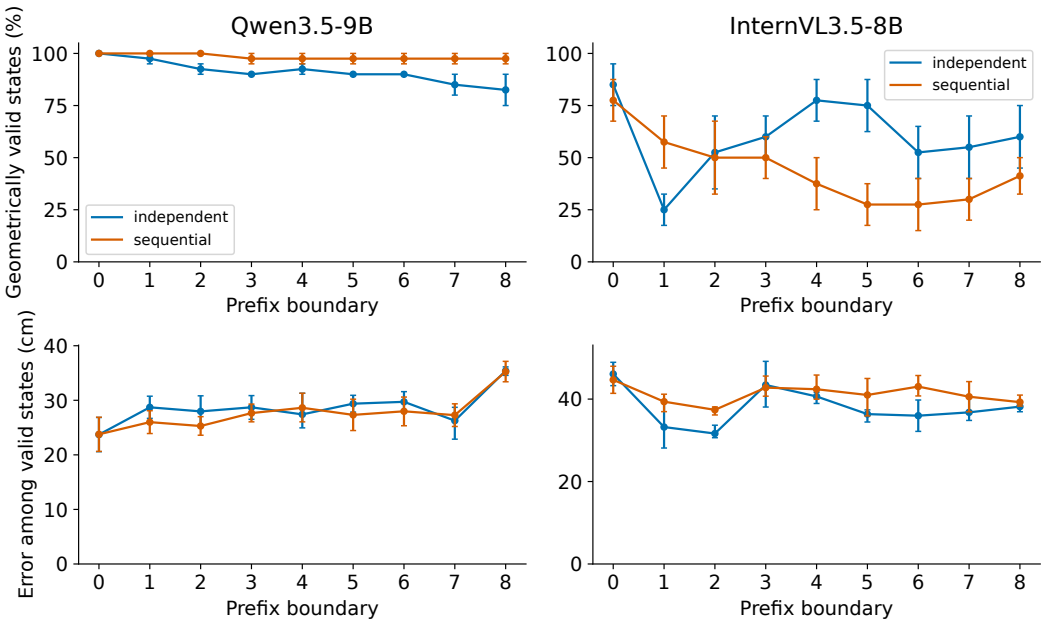


Figure 24: Valid JSON and valid geometry are separate requirements. Upper panels show geometrically valid states under the guided contract; lower panels average named-point distance only among valid outputs. The $T = 8$ cohort is fixed, but the contributing roots and task cells can change across boundaries. Missing valid states leave a gap. These continuous errors must be interpreted with unconditional coordinate success. CSV distances are meters and plotted distances are centimeters.

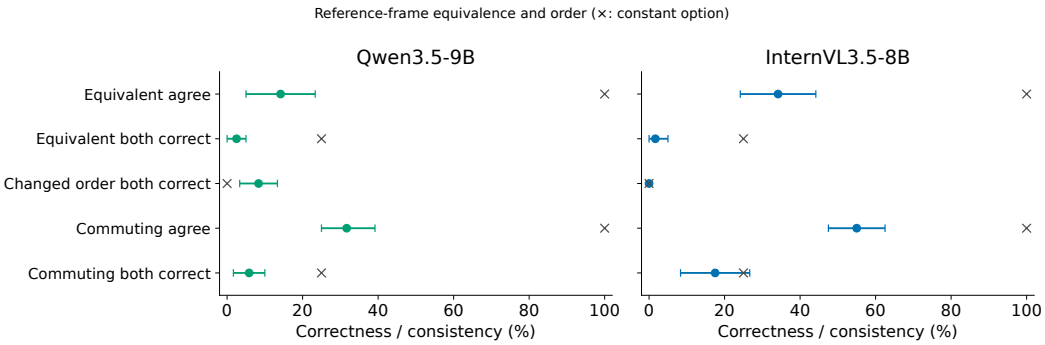


Figure 25: Equivalent programs are jointly correct at these rates: Qwen3.5-9B 2.50%; InternVL3.5-8B 1.67%. Gray crosses show the constant-option reference: 100% agreement, 25% jointly correct for unchanged answers, and zero jointly correct for changed-order answers. It is a trivial reference, not evidence of state tracking. Equivalent programs can agree without both being correct. The frame bank retains 44 questions in 11 source balancing groups (36 B, 8 E). Original and swapped programs have two actions and different verified answers; commuting controls split one action into two and compare equal-length three-action programs. Local-axis coefficients disclose numeric directional information, so equivalent-program differences are not a pure frame-language effect. Source-horizon stratification is retained; intervals are pointwise and some strata are small.

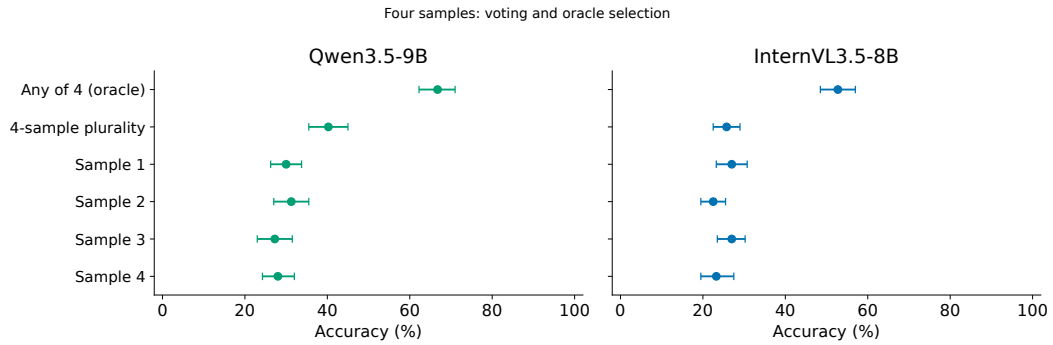


Figure 26: Oracle any-of-four accuracy bounds what an ideal selector could extract from these four recorded stochastic samples. Valid-answer plurality uses no ground-truth selection. Individual samples and the vote are shown on identical questions; the oracle result is not deployable accuracy.

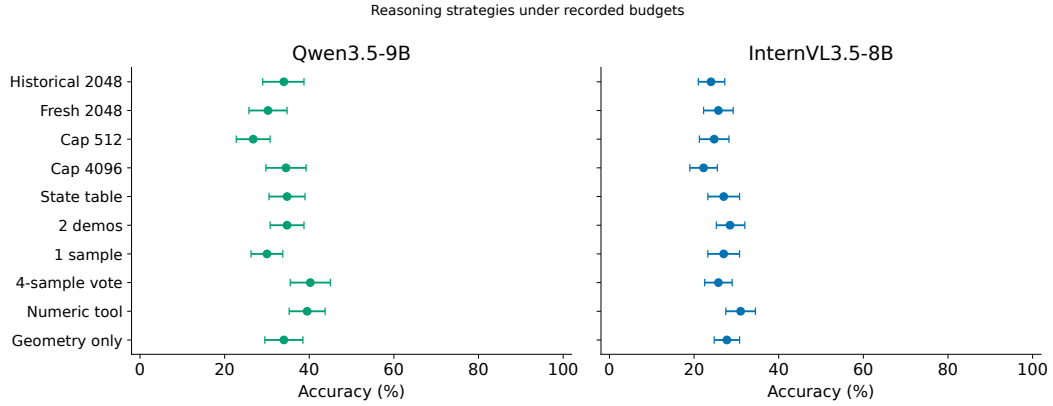


Figure 27: Four-sample voting accuracy is Qwen3.5-9B 40.25%; InternVL3.5-8B 25.75%. Reasoning aids are compared on the same 400 questions under declared budgets. Fresh baseline uses the same 2,048-token contract, seed, and serving concurrency as the follow-ups; the historical run is shown separately; caps change only the output limit, state tables and demonstrations change instructions, and four-sample voting uses four stochastic calls. Numeric tools operate on model-estimated initial state and require two model calls. Coordinate/tool prompts disclose static dimensions; geometry-only controls receive the same calibration. Latencies include local serving effects; no causal speed ranking is claimed. Intervals are pointwise 95%.

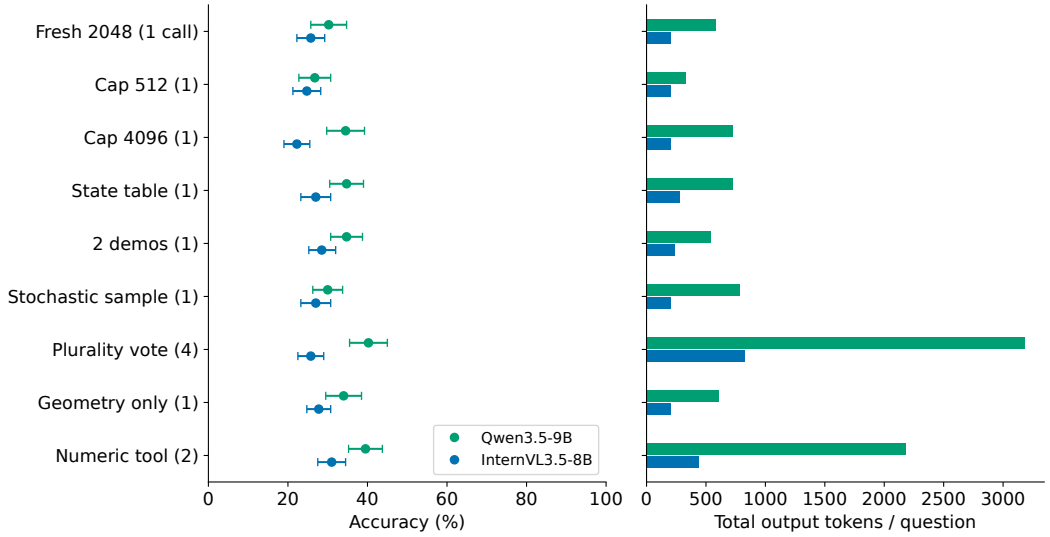


Figure 28: Accuracy is reported with the full output cost of each strategy. Call counts in labels are per question; voting includes four samples and tools include state extraction and final readout. Bars show all returned output tokens divided by 400 questions, with no uncertainty interval on this recorded accounting total. Accuracy intervals are pointwise 95%. Geometry is disclosed to the calibration and tool conditions. Token counts depend on tokenization and omit input computation and separately unmeasured tool time; they are not FLOPs or end-to-end latency.

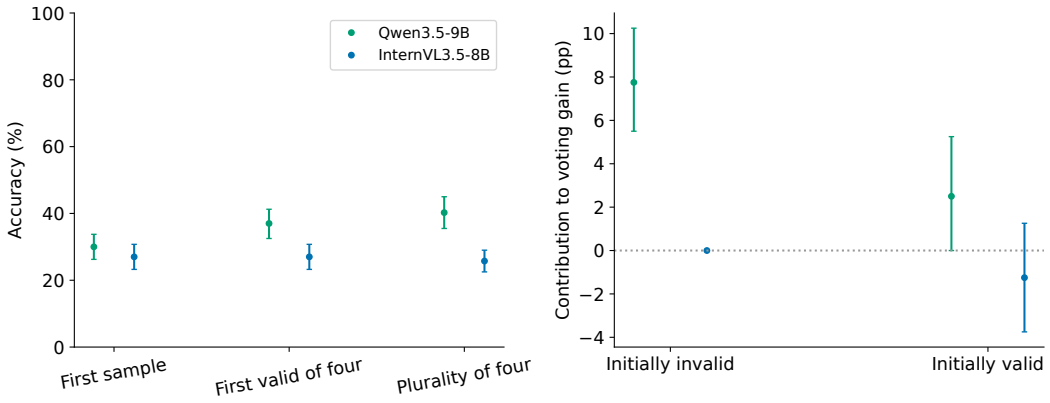


Figure 29: A first-valid selection control tests whether plurality adds benefit beyond obtaining a complete answer. This post-hoc analysis uses the four already recorded stochastic samples, retaining their order. Right: contributions to vote-minus-first-sample accuracy from questions whose first answer was invalid or valid; both use all 400 questions as denominator and sum to the total effect. The strata are descriptive. All four calls were made, so this is not an early-stopping cost experiment. Intervals are pointwise 95%.

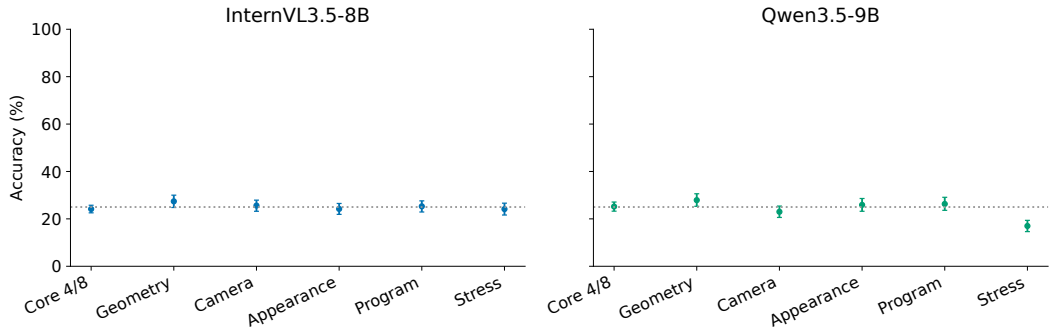


Figure 30: Accuracy remains limited across independently generated distribution-shift banks. Core reference uses horizons 4 and 8; each OOD bank has 1,000 questions and stress has 800. These are distribution comparisons, not within-scene causal rerender effects. Error bars are pointwise 95% balancing-group intervals; auxiliary scores are excluded from core accuracy.

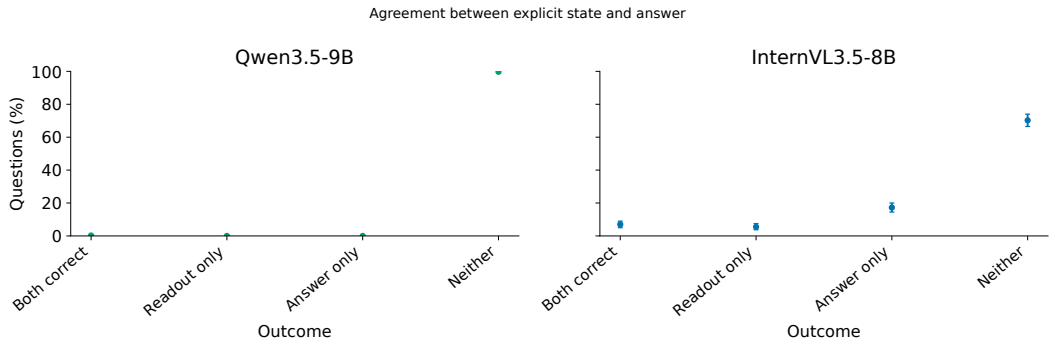


Figure 31: Original terminal-state queries conflate output completeness with coordinate readout. This panel compares a fixed readout of predicted coordinates and the separately emitted answer on 400 calibrated questions; missing outputs fail. Consult the original state-output completeness counts and separate guided-contract diagnostic before inferring weak state-answer coupling. This is explicit numeric output, not native image generation.

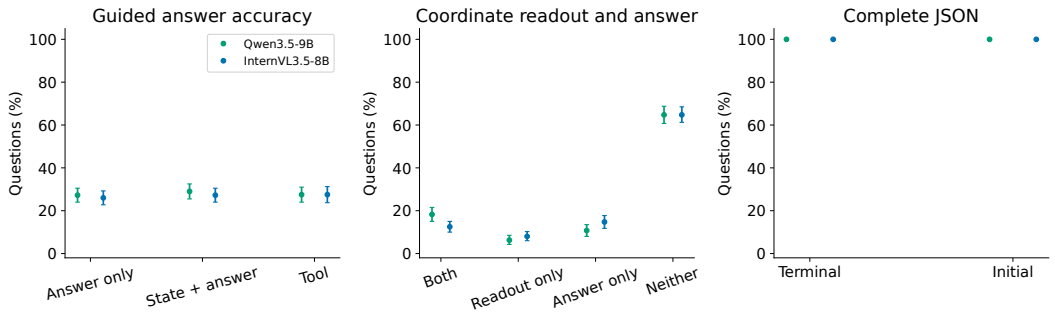


Figure 32: Schema-constrained outputs separate response completeness from geometric correctness. This post-hoc diagnostic changes both the immediate-JSON instruction and decoding, uses the same 400 calibrated questions, and preserves the original runs. Left: guided answer-only, joint state-and-answer, and estimated-state tool accuracy. Middle: fixed coordinate readout versus the jointly emitted answer. Right: complete JSON, which does not imply a valid rotation or accurate points. The answer-only control receives the same geometry and answer schema. All assigned questions remain in the denominator; intervals are pointwise 95%.

Condition / metric	Qwen3.5-9B	InternVL3.5-8B
answer_only / accuracy	27.25 [24.00, 30.50]	26.00 [22.75, 29.25]
terminal_state / accuracy	29.00 [25.50, 32.50]	27.25 [24.00, 30.50]
tool / accuracy	27.50 [24.00, 31.00]	27.50 [23.75, 31.25]

Table 69: Post-hoc guided JSON: strategies. All 400 assigned questions remain in the denominator. Values are percentages, or percentage points for differences, with pointwise 95% intervals. The calibrated answer-only control shares the answer schema and geometry disclosure.

Condition / metric	Qwen3.5-9B	InternVL3.5-8B
answer_only / valid_json	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]
answer_only / truncated	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
terminal_state / valid_json	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]
terminal_state / truncated	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
tool_initial / valid_json	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]
tool_initial / truncated	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
tool / valid_json	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]
tool / truncated	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

Table 70: Post-hoc guided JSON: format. All 400 assigned questions remain in the denominator. Values are percentages, or percentage points for differences, with pointwise 95% intervals. The calibrated answer-only control shares the answer schema and geometry disclosure.

I.10 Separate guided-contract results

Condition / metric	Qwen3.5-9B	InternVL3.5-8B
tool_failure / fraction	19.75 [17.00, 22.75]	41.75 [38.00, 45.50]
executor_accuracy / fraction	27.75 [24.25, 31.25]	12.25 [9.75, 15.00]
consistent / fraction	59.75 [54.75, 64.50]	51.25 [46.00, 56.50]

Table 71: Post-hoc guided JSON: diagnostics. All 400 assigned questions remain in the denominator. Values are percentages, or percentage points for differences, with pointwise 95% intervals. The calibrated answer-only control shares the answer schema and geometry disclosure.

Condition / metric	Qwen3.5-9B	InternVL3.5-8B
tool_minus_answer_only / difference	0.25 [-3.25, 3.75]	1.50 [-0.75, 3.75]

Table 72: Post-hoc guided JSON: differences. All 400 assigned questions remain in the denominator. Values are percentages, or percentage points for differences, with pointwise 95% intervals. The calibrated answer-only control shares the answer schema and geometry disclosure.

Condition / metric	Qwen3.5-9B	InternVL3.5-8B
Both correct / fraction	18.25 [15.00, 21.51]	12.50 [10.00, 15.00]
State only / fraction	6.25 [4.25, 8.50]	8.00 [6.00, 10.25]
Answer only / fraction	10.75 [8.00, 13.50]	14.75 [11.75, 17.75]
Neither / fraction	64.75 [60.75, 68.75]	64.75 [61.25, 68.50]

Table 73: Post-hoc guided JSON: coupling. All 400 assigned questions remain in the denominator. Values are percentages, or percentage points for differences, with pointwise 95% intervals. The calibrated answer-only control shares the answer schema and geometry disclosure. State correctness means the coarse coordinate-derived option, not full pose accuracy.

Quantity	Qwen3.5-9B	InternVL3.5-8B
first_sample	30.00 [26.25, 33.75]	27.00 [23.25, 30.75]
first_valid_of_four	37.00 [32.50, 41.25]	27.00 [23.25, 30.75]
vote_four	40.25 [35.50, 45.00]	25.75 [22.50, 29.00]
first_valid_minus_single	7.00 [4.75, 9.25]	0.00 [0.00, 0.00]
vote_minus_first_valid	3.25 [0.75, 6.00]	-1.25 [-3.75, 1.26]
gain_on_initially_invalid	7.75 [5.50, 10.25]	0.00 [0.00, 0.00]
gain_on_initially_valid	2.50 [-0.00, 5.25]	-1.25 [-3.75, 1.26]

Table 74: Post-hoc voting completeness control. Accuracy is in percent and differences/contributions are percentage points. Contributions from initially valid and invalid responses use all 400 questions and sum to the total voting gain. All four samples were executed.

Model	Original state query	Complete JSON	Length-limited	Total
Qwen3.5-9B	terminal_state	1	399	400
Qwen3.5-9B	tool_initial	51	285	400
InternVL3.5-8B	terminal_state	298	0	400
InternVL3.5-8B	tool_initial	364	0	400

Table 75: Original coordinate-output completeness. A decoded state object still needs geometric validation. Length termination and JSON completeness are separately counted and need not be complements.

J CONSTRUCTION SCHEMATIC AND ACTION REPLAYS

Figure 33 follows one verified pair from sampling to scoring; Figure 34 shows the action boundaries for three development examples. The replays expose hidden identities for review, so they illustrate label computation rather than the primary model input.

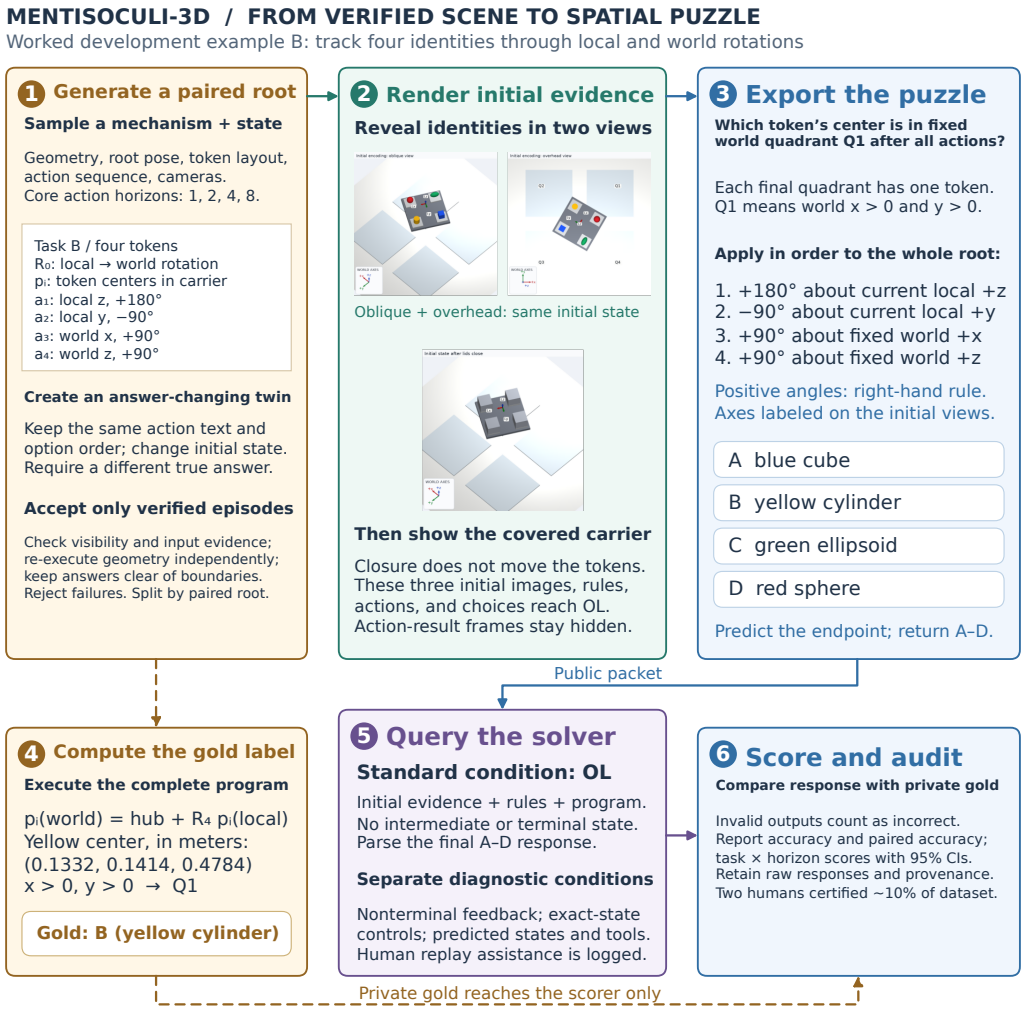


Figure 33: **A verified initial scene and action program become a four-choice spatial puzzle.** The worked development example follows task B from paired-root generation to scoring. The three initial images reveal identities before closure; the closure moves no token. Solid arrows carry the construction output or public model packet; dashed arrows carry private execution and gold-label information. The numerical readout identifies the yellow cylinder in Q1 after four rotations. Acceptance checks reduce ambiguity but do not establish human readability.

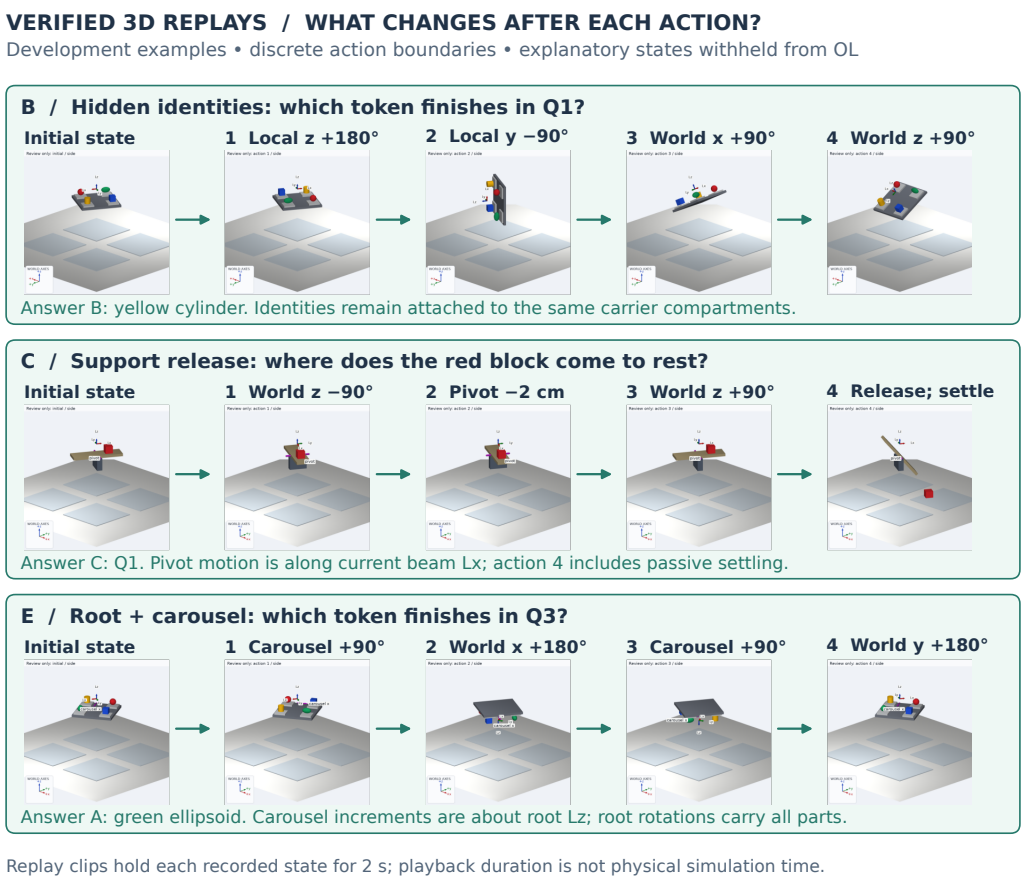


Figure 34: **Stored action boundaries expose the transformations that a solver must predict.** Each row shows the initial state and all four executed action endpoints for one development example: hidden identities (B), support release (C), and root plus carousel rotation (E). Identity-revealing review views expose covered tokens for explanation; these result frames are withheld from the standard OL condition. The rightmost C frame shows the accepted settled state, not an instantaneous release. World axes are fixed; local axes follow their current parent frame. Positive rotation follows the right-hand rule, appearing anticlockwise when viewed from the positive-axis tip toward the origin.

The companion `schematic.html` provides the editable diagrams, vector exports, and three replay clips. Each clip holds a stored state for two seconds; this presentation interval is neither physical simulation time nor a continuous motion reconstruction. The source assets contain only these worked development examples. Their recorded programs, answer options, numeric states, and frame hashes are retained with the figures.

K COMPLETED FULL-CORE EXTENSION AND ASTRA PROTOCOL

K.1 COVERAGE AND INFERENCE SETTINGS

The full-core extension completes Qwen3.8-27B-FP8 and adds GPT-6 Astra on the same frozen 4,000 OL questions. Neither run changes the bank, labels, option order, or answer parser. No question is selected or discarded based on a model's answer. The original matched-400 and intervention results remain separate; Astra has not been evaluated on those intervention or distribution-shift contracts.

The 27B run uses revision `017b9c7af6b5689d5dd426a76e0bc077eb5ca20a`, official FP8 weights, vLLM 0.29.0, four concurrent requests, an 8,192-token context, and a 1.5 GiB KV-cache allocation. Its image processor, temperature zero, 2,048-token output cap, brief instruction, and template-level thinking setting match the recorded open-model protocol. Cache capacity was selected through development-only execution checks; a 2 GiB attempt exhausted image-encoder memory before core scoring. The new full-core run imports no responses from the earlier matched run or the separate, incomplete CPU 4-bit run. All 4,000 responses are retained across two resumable GPU windows.

Astra uses the requested alias `gpt-6-astra` in Codex CLI 0.156.1, authenticated through an existing ChatGPT login, on 25 September 2026. Each question starts a new `codex exec --ephemeral` process with `model_reasoning_effort="high"`. At most four independent sessions run concurrently. The CLI supplies no immutable provider snapshot identifier, so the recorded alias, version, command, timestamps, and response events define the available provenance; bitwise reproducibility is not claimed.

Each request contains only the public rule card, initial images, numbered image captions, action program, question, and four options. Original image bytes are attached in their original order through `--image`. The captions are grouped in text rather than interleaved through the local API. The brief instruction requests at most 150 visible words and one terminal `FINAL: X` tag. Native reasoning is set to high; no 2,048-token cap, temperature, or sampling seed is enforced. Provider image processing is not assumed equivalent to the local models' processor settings. These differences preclude a compute-matched causal interpretation of the score gap.

K.2 ISOLATION, RECORDS, AND SCORING

A fresh temporary directory contains only the current images and a fixed instruction file. User configuration and project documentation are excluded; browsing, shell execution, apps, plugins, delegation, and host skill discovery are disabled or excluded. A filesystem permission profile denies tool reads under the project user's home directory. The model is instructed to use only the supplied puzzle. Any observed tool activity makes an execution ineligible and stops new admissions. No tools were observed in the 4,000 retained core event streams. Answer keys, simulator states, and prior model responses were not included in requests.

Ten fixed development questions, two per task and including two- and three-image inputs, checked execution before test inference. The gate required one completed turn, agreement between the final-message file and event stream, and no tool activity or unexpected errors. It did not use development correctness to select settings. Compressed requests preserve text and images; raw records preserve commands, final text, CLI events, usage, timing, and packet/image hashes. Usage-limit or infrastructure errors stop admission instead of triggering credential changes or answer-dependent retries. The frozen parser scores invalid model answers as incorrect. Simulator labels are used after inference for scoring; Astra is a tested solver, not a replacement answer-key judge.

All 8,000 new requests and corresponding raw records passed packet, image, request-hash, and parser checks. The Astra event audit found zero tool calls, and both runs have zero infrastructure errors. Independently rebuilding the overall, pair, task, and horizon confidence intervals reproduced all 22 reported intervals for these two runs from their public anonymized score vectors. Table 76 shows that Astra's high accuracy covers every task family, whereas the 27B run performs best on support release. Table 77 shows that Astra also retains high accuracy at eight actions. These marginal profiles average independent questions and do not isolate the effect of increasing the length of one program.

Model	A	B	C	D	E
Qwen3.8-27B FP8	21.63	23.63	65.50	18.00	29.38
GPT-6 Astra (high)	98.25	99.25	99.00	97.00	99.63

Table 76: Full-core task accuracy in percent for the two new runs; each task contains 800 questions. Contracts differ as specified in Table 54.

Model	$T = 1$	$T = 2$	$T = 4$	$T = 8$
Qwen3.8-27B FP8	56.30	38.20	21.40	10.60
GPT-6 Astra (high)	99.00	99.30	97.70	98.50

Table 77: Full-core accuracy in percent by action horizon, with 1,000 independently sampled questions per horizon. These are not extensions of the same scenes.

K.3 Interpretation and reproduction

Astra answers 3,945 of 4,000 questions correctly, leaving 55 errors. Its accuracy ranges from 97.00% to 99.63% across task families and is 98.50% at eight actions. High pair accuracy also rules out solving these pairs through a deterministic shared-text-only answer rule. These results establish high performance on the tested finite distribution; they do not identify an internal mental-imagery mechanism or establish human readability. The near-ceiling result limits the benchmark's resolution among stronger systems.

```
python scripts/audit_recent_core.py
python scripts/write_recent_paper_results.py
# Local retained requests are required for this additional check:
python scripts/audit_recent_core.py --private
# New inference requires the prepared bank and authorized Codex login:
python scripts/run_astra_core.py --concurrency 4
```

The generated comparison table reports all eight completed core evaluations and separates the Astra inference contract. The public source manifest records the score-summary and generated-output hashes. Raw per-question answers remain in the private evaluation archive.