TENONBENCH: EVALUATING SPATIOTEMPORAL CAPABILITIES WITH MORTISE-AND-TENON ASSEMBLIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating whether vision-language models (VLMs) can reason about feasible motions and operation order is essential for their use in embodied planning. However, existing benchmarks provide limited coverage of changing geometric constraints, do not fully distinguish necessary from optional ordering, and often assess perception and planning separately. To address these gaps, we introduce **TenonBench**, a spatiotemporal reasoning benchmark that combines interlocking assembly tasks, multi-solution evaluation, and paired perception–reasoning analysis. We construct 178 parametric mortise-and-tenon assemblies with 1,200 parts and use a geometric solver to provide ground truth for 4,065 questions spanning part identification, structural understanding, and reasoning about how movement direction and preceding removals determine motion feasibility. Our partial-order evaluation distinguishes necessary dependencies from optional ordering that varies across valid solutions, while complete sequences are verified by replaying each removal against the remaining assembly. Within multi-turn shared conversations, we further pair reasoning questions with perception probes about the same parts and jointly analyze their responses to examine whether reasoning errors persist when the corresponding perception questions are answered correctly. Evaluation of 23 VLMs yields a top accuracy of 77.2% and a mean of 53.7%, highlighting TenonBench as a challenging testbed for geometry-grounded spatiotemporal reasoning.

1 INTRODUCTION

Recent advances in visual understanding and instruction following have made vision-language models (VLMs) promising foundations for embodied agents, with applications ranging from household assistance to robotic assembly and industrial maintenance (Yang et al., 2025). Translating visual observations into feasible physical operations, however, requires more than recognizing objects and their relative positions. Models must infer three-dimensional geometric constraints from images and understand how these constraints govern motion. In rigid-part assemblies, a component may be blocked in one direction but removable in another, while removing one part can unlock the motion of others. **Motion feasibility and operation order** are therefore tightly coupled. Assessing whether VLMs understand this coupling is an important step toward evaluating their ability to reason about physical operations.

Existing benchmarks assess spatial and physical intelligence through spatial relations and transformations (Yang et al., 2026; Zhang et al., 2025b), physical interaction prediction (Chow et al., 2025), and assembly understanding and planning (Wang et al., 2022; Ma et al., 2026a; Chetan et al., 2026). Despite this progress, three gaps remain in the evaluation of coupled rigid-part operations. First, these benchmarks give limited attention to how motion feasibility depends jointly on movement direction and the parts still in place, and how earlier operations change collision constraints to enable later actions. Second, sequence correctness alone does not establish whether a model understands which dependencies are necessary. Matching a single reference can reject valid alternatives, whereas accepting any feasible sequence does not reveal whether the model invents unnecessary dependencies or overlooks required ones. Third, perception and planning are often evaluated separately, without matching questions about the same parts. This limits analysis of whether operation-reasoning errors coincide with errors on related structural questions or persist when those questions are answered correctly.

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.

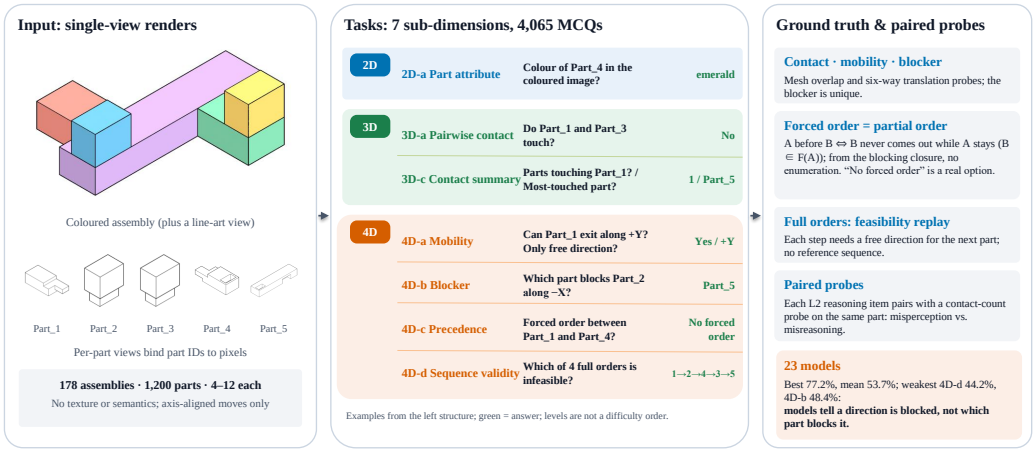

Figure 1: **Overview of TenonBench.** Left: the input of each assembly is a colored whole-assembly render (plus a line-art render) and per-part views. Middle: one example question per sub-dimension, all from the 5-part structure on the left; green marks the reference answer. Right: all ground truth comes from the geometric solver; pairwise precedence is a partial order computed from the blocking closure, and every L2 reasoning question is paired with a perception probe on the same part.

To address these gaps, we introduce **TenonBench**, a benchmark for *spatiotemporal reasoning* that evaluates whether VLMs can infer feasible motions and necessary operation dependencies from visual observations (Figure 1). TenonBench uses interlocking mortise-and-tenon assemblies, in which the feasibility of removing a part depends on both its movement direction and the parts that remain. These structures provide a controlled setting for examining how geometric constraints change as operations proceed. Under a motion model restricted to axis-aligned translations, a geometric solver computes labels for contact, directional mobility, blocking relations, and operation dependencies. TenonBench combines three complementary evaluation designs. First, its motion-reasoning tasks assess how direction-specific blocking and preceding removals determine whether an operation is feasible. Second, its partial-order evaluation distinguishes necessary precedence from relative orders that vary across valid solutions, while complete sequences are checked through step-by-step feasibility replay. Together, these evaluations accommodate multiple valid solutions and separately expose invented and missed dependencies. Third, questions progress from perception to planning within a shared conversation, with part-specific reasoning questions paired with perception probes about the same parts. Joint response analysis measures whether reasoning errors persist when the matched perception probes are answered correctly, without assuming that success on a probe establishes complete geometric understanding.

Specifically, TenonBench comprises 178 parametric assemblies with 4–12 parts each, totaling 1,200 parts and 4,065 multiple-choice questions. Each assembly is presented through colored and line-art whole-assembly renders, together with per-part views that establish part identities. Questions span three capability dimensions: 2D part identification, 3D static structure, and 4D motion constraints and dependencies, where the temporal component concerns operation order and the resulting changes in assembly state. They are further organized into three scopes: a named part or pair (L1), a part in relation to the rest of the assembly (L2), and the whole assembly (L3). Part count and disassembly depth provide additional axes for analyzing difficulty. Every question includes an “unsure” option, and all 1,470 L2 reasoning questions are paired with contact-count probes for the corresponding parts. We evaluate 23 VLMs using the same prompts and scoring protocol. The best model achieves 77.2% overall accuracy, while the mean across models is 53.7%. Beyond aggregate performance, the results reveal persistent difficulties in attributing motion constraints to specific blockers and distinguishing necessary dependencies from optional ordering. Reasoning errors also remain when the matched perception probes are answered correctly, indicating that success on local structural questions does not ensure correct reasoning about physical operations.

In summary, our contributions are:

- We introduce TenonBench, a benchmark comprising 178 mortise-and-tenon assemblies, 1,200 parts, and 4,065 multiple-choice questions. Its tasks cover three capability dimensions—2D part identification, 3D static structure, and 4D motion constraints and dependencies—and three ques-

tion scopes: a named part or pair (L1), a part in relation to the remaining assembly (L2), and the whole assembly (L3).

- We develop a geometric solver that computes ground truth for contact, directional mobility, blocking relations, and disassembly dependencies. It derives necessary pairwise precedence from the blocking closure without enumerating valid sequences, distinguishing required precedence from optional ordering. Complete disassembly sequences are verified through step-by-step feasibility replay rather than comparison with a single reference sequence.
- We provide an evaluation framework covering perception and operation reasoning, with separate metrics for invented and missed dependencies and accuracy on full-sequence validity questions. All 1,470 L2 reasoning questions are paired with contact-count probes about the same parts in shared conversations. Joint analysis reports probe–reasoning correctness and reasoning accuracy conditional on probe correctness.
- We evaluate 23 VLMs, with a best overall accuracy of 77.2% and a mean of 53.7%. The results show errors in both identifying necessary dependencies and recognizing optional ordering. Reasoning errors persist even when matched contact-count probes are answered correctly, while performance also varies with question scope and assembly part count.

Table 1: **Task system.** Scopes L1–L3 are defined in Section 3.1; example questions per type are in Appendix M.

Dim.	Task	Core question	Scope	#Q
2D-a	Part attribute	Color of a named part in the colored assembly	L1	356
3D-a	Pairwise contact	Whether two parts touch in the assembled state	L1	534
3D-c	Contact aggregation	How many parts a named part touches; which part touches most	L2, L3	1,351
4D-a	Directional mobility	Whether a named part can be removed alone along a given direction	L2	878
4D-b	Blocker attribution	Which single part blocks it along that direction	L2	356
4D-c	Disassembly dependency	Whether two parts have a forced precedence; which part can be removed first	L2, L3	413
4D-d	Full-sequence validity	Which of four full disassembly orders is infeasible	L3	177

2 Related Work

Assembly and block benchmarks. Assembly benchmarks provide a natural setting for evaluating multi-step spatial reasoning. LEGO-Puzzles (Tang et al., 2025) and PhyBlock (Ma et al., 2026a) use LEGO construction and 3D block assembly, while furniture benchmarks assess manual understanding and assembly-step ordering (Wang et al., 2022; Tie et al., 2025; Chetan et al., 2026). Related work studies assembly actions and errors in videos (Ben-Shabat et al., 2021; Sener et al., 2022; Ding et al., 2023), fragment reassembly (Sellán et al., 2022), and task-oriented assembly assistance (Jing et al., 2026; Huang et al., 2025). These settings emphasize construction procedures, support relations, or observed sequences, with less focus on distinguishing necessary dependencies from optional ordering under interlocking constraints. TenonBench complements them with assemblies whose motion feasibility depends jointly on motion direction and preceding removals. It evaluates necessary pairwise precedence alongside full-sequence feasibility, distinguishing necessary dependencies from optional ordering.

Spatial-reasoning benchmarks. Spatial benchmarks span basic relations, multi-view understanding, mental transformations, and interactive or causal reasoning (Kamath et al., 2023; Yang et al., 2026; Huang et al., 2026; Wasi et al., 2026; Zhang et al., 2025b; Mayer et al., 2025; Ma et al., 2026b). Some also investigate model failures through error annotations (Yang et al., 2026; Huang et al., 2026), dedicated perception questions (Khezresmaeilzadeh et al., 2026), or text-mediated separation of perception and reasoning (Qiao et al., 2024). TenonBench focuses on a complementary question: whether a motion is geometrically feasible and how executing it changes the feasibility of later operations. It also pairs L2 reasoning questions with contact-count probes about the same parts within shared conversations, enabling analysis of reasoning performance conditioned on matched perception-probe correctness.

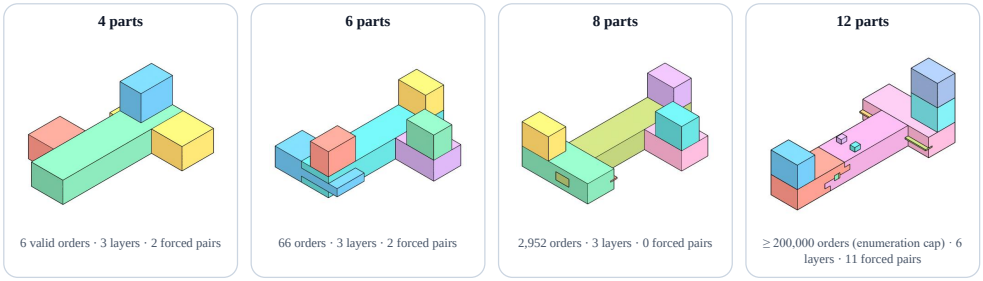

Figure 2: One assembly at each of four sizes (colored renders), with its number of valid disassembly orders, disassembly layers and forced part pairs. The 12-part structure exceeds the enumeration cap, and its exact count comes from the closure and the dynamic program.

Physical understanding and embodied QA. PhysBench (Chow et al., 2025) and Physion (Bear et al., 2021) evaluate physical properties, interactions, and outcome prediction, while counterfactual and minimal-pair designs help assess reliance on shortcuts (Baradel et al., 2019; Krojer et al., 2025). Embodied benchmarks cover situated question answering (Ma et al., 2022; Majumdar et al., 2024), skill composition (Haresh et al., 2024), agent task execution (Yang et al., 2025), and interaction or manipulation reasoning (Wang et al., 2026; Li et al., 2026). TenonBench complements these broader evaluations by isolating rigid geometric interference under an axis-aligned translation model, excluding gravity, friction, and deformation. Its solver-derived labels support separate evaluation of directional mobility, blocker attribution, and operation dependencies, providing a focused assessment of geometric constraints underlying physical operations.

3 The Benchmark

3.1 Overview and Task System

Answering needs three kinds of information, and the tasks form three capability dimensions, 2D, 3D and 4D, with seven sub-dimensions (Table 1): **2D part identification** links the part ID in the question to the per-part view and the colored assembly—the input to the other two dimensions; **3D static structure** asks which parts touch, how many parts each touches and which is most connected, without motion; **4D motion constraints and dependencies** asks about motion feasibility—whether a part can be removed along a direction and which part blocks it, and, as parts are removed and the state changes, pairwise forced precedence and the validity of full disassembly sequences. The tags are organizational: "time" in 4D means operation order and state change while the input stays static; the capability dimensions are not a difficulty ranking (Section 4.2).

Orthogonal to the dimensions, each question has a *scope*, the information needed to answer it. **L1** involves only the part or pair named in the question and is read directly from the images (color, pairwise contact). **L2** relates a named object to the rest of the structure, which must first be located within the whole (contact count, directional mobility, blocker attribution, pairwise precedence). **L3** concerns the whole structure without naming a part and requires integrating all parts (most-connected part, first removable part, invalid sequence identification). Scope and dimension do not coincide: within 3D-c, contact counting is L2 and the most-connected question L3. There are 890 L1, 2,670 L2 and 505 L3 questions; the perception–reasoning pairing of Section 3.2 is made on L2.

Assemblies have 6.74 parts on average (Figure 2); the median number of disassembly layers is 3 and 73.0% of assemblies need three or more, so cross-layer dependency is the norm. The 4,065 questions span ten types over the seven sub-dimensions; counts are unbalanced (color questions are limited by the coverage of the colored render), so results are reported per type (Appendix B); four representative questions on one structure are in Appendix Figure 18, and set statistics in Appendix J.

3.2 Benchmark Construction

Assets and visual input. Basic mortise-and-tenon units, each a complete joint with a few parts, are combined in all unordered pairs into larger interlocked assemblies, giving 498. Each is exported as a multi-part mesh and single-view renders: a line-art whole-assembly render, a whole-assembly render in a fixed palette, and a per-part view of every part, which binds part IDs to the image.

Parts are untextured cuboid blocks (M3); residual priors are measured in Appendix F and control conditions designed in Appendix I. Only single fixed-view renders are evaluated.

Question instantiation. Questions are instantiated from the templates of Table 1 with labels from Section 3.3. Each type is generated only when the answer is unique: no most-connected question on tied degrees, unique-free-direction questions only when exactly one direction is free, blocker questions only when the blocker is unique, and color questions only when a visibility guard confirms the part is visible (Appendix K). Every question includes an "unsure" option so that abstention and error are separable.

Multi-solution tasks. The pairwise task offers, besides the two directions, the explicit option "no forced order between them; either order works"; the full-sequence task offers three valid orders and one infeasible one, preventing models from relying on simple surface-level heuristics. Forced and free pairs are balanced over the whole set: only 108 of 3,699 pairs (2.9%) are forced and 127 assemblies contain none, so natural-frequency sampling would push the majority baseline near 100%; set-level sampling gives 132 free and 104 forced items (majority baseline 55.9%).

Perception probes and filtering. Every part gets one contact-count question as a perception probe (1,200 in total); each of the 1,470 L2 reasoning questions (mobility, blocker, pairwise precedence) is paired with the probe on the same part (782 parts carry at least one). Of the 498 assemblies, 465 have a deadlock-free disassembly sequence; the evaluation set keeps the 178 whose geometric decisions are in a reliable regime (criteria in Appendix K); failing assemblies are dropped whole.

3.3 Geometric Ground Truth and Verification

Local labels. All ground truth is computed on the 3D assets; models see only the renders. Contact is decided by translating each of two parts a small step along each of the six directions, intersecting the meshes and thresholding the overlap volume; contact counts and the most-connected part are read from the contact graph. Directional mobility, the single-step motion feasibility of a part along a direction, is decided by translating the part a fixed fraction of its extent and testing whether its total overlap with the other parts grows; the blocker is the part whose overlap grows.

Forced precedence without enumeration. Take the state to be the set S of remaining parts: p is removable in S iff some direction has all its blockers outside S. A must precede B iff no valid order places B before A. This is decidable without enumeration: keep A in place and repeatedly remove any other removable part until none is left; call the remaining set $F(A)$. Then A must precede B is equivalent to $B \in F(A)$. Removing a removable part only shrinks the blocking sets of the others, so the order of removal does not affect the terminal state and $F(A)$ is unique: it is the set of parts that cannot be removed while A stays, and $B \in F(A)$ means every order placing B before A gets stuck before B. A 4-part example is in Appendix Figure 16. Parts removable first are those with a free direction initially; the number of valid orders is counted by dynamic programming over subsets.

Replay and cross-validation. Every candidate order in a full-sequence question is replayed step by step, as in plan validation (Valmeekam et al., 2023); the order is infeasible at the first step whose part has no feasible direction, and the failing step and its blockers are recorded. The closure solver agrees with brute-force enumeration on all 169 assemblies where enumeration completed, and 5 assemblies were reassembled by a robot arm in simulation following the solver's order (Appendices K and L).

4 Experiments

4.1 Experimental Setup

We evaluate 23 VLMs: 13 closed API models (4 GPT, 4 Gemini, 3 Claude, 2 commercial Qwen) and 10 open-weight models (3 Qwen, 2 Kimi, 2 InternVL (Wang et al., 2025), 2 Cosmos (Agarwal et al., 2026), BAGEL (Deng et al., 2025)). Models answer all 4,065 questions with the same English prompts and scoring program. All questions of one assembly are asked in one shared conversation (178 conversations); images are given once in the first turn, questions are ordered by increasing dimension, and the paired probe precedes its reasoning question. The model gives a one-sentence rationale grounded in images, then chooses an option. Failed calls and unparsable outputs are excluded from the denominator; "unsure" counts as wrong and is also reported as abstention (Appendix A).

Table 2: **Accuracy (%) per dimension.** Every model answered all 4,065 questions; dark gray (bold) marks the best and light gray the second best model in each column. Models are grouped by family and ordered by overall accuracy. The last two rows are the random and majority-class baselines.

Model	Overall	2D-a	3D-a	3D-c	4D-a	4D-b	4D-c	4D-d
Closed API								
gpt-6-astra	**77.2**	**99.2**	**86.1**	65.4	**72.4**	**84.8**	**86.7**	**81.9**
gpt-5.6-terra	67.8	69.4	77.7	64.2	63.3	64.0	77.7	67.8
gpt-5-mini	47.5	49.4	57.3	42.2	46.8	49.2	52.8	39.3
gpt-5.4-nano	34.7	26.1	46.3	36.4	29.1	33.7	39.0	23.7
gemini-3.8-flash	76.7	88.1	79.5	**74.7**	70.1	76.3	82.3	80.6
gemini-3.1-pro-preview	72.7	90.3	80.4	67.6	68.1	66.7	81.6	65.9
gemini-3.5-flash	72.3	90.2	83.7	64.6	70.2	69.7	77.0	65.7
gemini-3.1-flash-lite	60.6	59.6	77.2	55.2	64.0	48.9	65.4	50.3
claude-opus-5	71.4	81.4	83.4	65.1	65.3	72.4	78.9	73.4
claude-sonnet-4-6	53.7	32.6	65.0	53.4	58.5	45.2	61.5	39.0
claude-haiku-4-5	43.0	31.2	50.0	40.3	49.9	37.1	49.2	30.5
qwen3.8-max	64.5	73.0	76.4	56.0	65.7	61.5	71.2	58.9
qwen3.8-flash	57.9	64.3	64.5	54.4	60.0	46.2	63.9	50.0
Open weights								
kimi-k3	58.1	66.6	70.8	50.2	60.9	50.1	65.4	47.5
kimi-k2.5	53.0	50.8	71.0	46.4	56.7	39.9	59.3	46.8
Qwen3.8-27B	56.6	59.3	67.2	50.6	59.8	47.8	68.3	40.1
Qwen3.5-35B-A3B	41.3	31.5	50.4	42.0	42.4	30.4	47.6	28.2
Qwen3.5-27B	37.8	33.4	49.4	35.9	42.0	26.4	41.1	19.5
InternVL3.5-38B	49.5	48.3	60.9	46.0	56.5	45.8	49.2	17.5
InternVL3.5-8B	38.1	36.0	38.2	39.4	37.2	36.8	42.1	29.9
Cosmos3-Super	38.7	40.5	49.6	44.8	27.6	29.8	42.6	19.8
Cosmos3-Nano	32.6	33.2	46.4	31.5	32.6	18.8	37.5	14.1
BAGEL-7B-MoT	30.6	25.3	37.8	30.0	24.3	31.0	43.6	25.4
Mean of 23 models	53.7	55.6	63.9	50.3	53.2	48.4	60.2	44.2
Random baseline	23.4	20.3	33.3	20.0	25.2	20.0	24.7	20.0
Majority baseline	40.9	17.1	66.7	32.4	60.9	22.8	41.4	11.3

Table 2 ends with a random baseline (by option count, 23.4% overall) and a majority-class baseline, well above random where labels are unbalanced: 66.7% on contact and mobility, 55.9% on the balanced pairwise task.

4.2 Main Results

Per-dimension accuracy is in Table 2 (per type: Appendix Table 3). First, the benchmark is neither solved nor trivial: the best model reaches 77.2% and the weakest 30.6%, the mean of 23 models is 53.7%, all above the 23.4% random baseline.

Second,difficulty does not strictly track the 2D–4D progression; rather, the primary performance drop occurs between pairwise and aggregate judgments. The three lowest dimension means are 4D-d full-sequence validity (44.2%), 4D-b blocker attribution (48.4%) and 3D-c contact aggregation (50.3%); the two highest are 3D-a contact (63.9%) and 4D-c disassembly dependency (60.2%). The three lowest all aggregate several local relations into one conclusion—counting all contacts of a part, locating the blocker of one direction among six, replaying a whole order—whereas 3D-a and 4D-c need one pairwise judgment, in line with the multi-object limits of VLMs (Campbell et al., 2024; Rahmanzadehgervi et al., 2024). The best model exhibits a similar performance profile: 3D-c (65.4%) is its lowest dimension (the 4D-c mean is inflated by one-sided answering, Section 4.3). 4D-a and 4D-b ask about the same part and direction: 4D-a whether the part can be removed alone along, say, $-X$ (removable or blocked), 4D-b which single part blocks it there (one of the other parts). 15 of 23 models score higher on 4D-a (mean gap 4.8 points; Figure 3e): judging that a direction is blocked is on average easier than naming the blocking part.

Third, within model families, full-sequence validity is where scale and tier matter most. In the three open families (Figure 3a–c), 4D-d is the lowest dimension of all seven models and does not reliably improve with scale (InternVL3.5 gains 11.4 points overall from 8B to 38B while 4D-d falls from 29.9% to 17.5%). In the GPT, Gemini and Claude families, the best model scores above its own

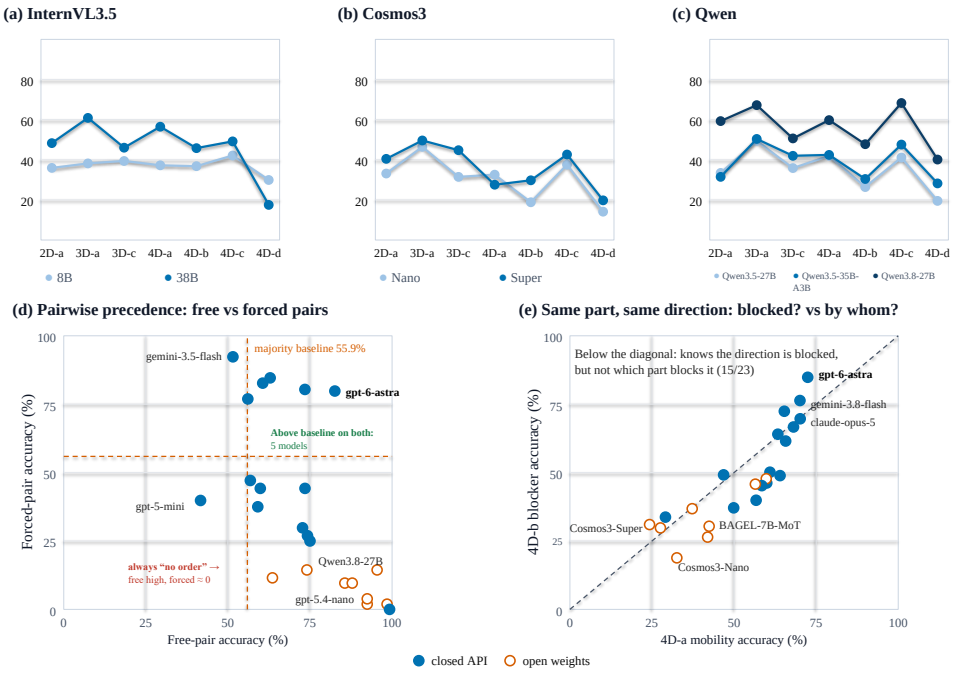

Figure 3: **Evaluation across sub-dimensions and pairwise behavioral diagnostics.** (a)–(c) Accuracy of three representative open-weight families (InternVL3.5, Cosmos3, Qwen) across the seven sub-dimensions; within each family, darker curves indicate larger or newer models. (d) Pairwise precedence accuracy split by ground truth, plotting free-pair accuracy against forced-pair accuracy; (e) 4D-a directional mobility vs. 4D-b blocker attribution for the same part and direction; the dashed diagonal represents parity, with 15 of 23 models falling below the diagonal.

overall accuracy on 4D-d and the smallest tier below it, so 4D-d spans 30–58 points within a family, more than overall accuracy (16–42; Table 2).

4.3 Two Kinds of Precedence Error

Pairwise precedence (4D-c) asks for the necessary dependency between two parts. A question names two parts and offers "A must be removed before B", "B before A", "no forced order; either order works" and "unsure". A pair is *forced* if one part precedes the other in every valid order (Section 3.3), and the answer is that direction; it is *free* if both orders occur in some valid order, and the answer is "no forced order". A forced pair requires seeing the blocking dependency, a free pair recognizing that none exists. We call the share of free pairs answered with an order the *invented-dependency rate*, and the share of forced pairs answered "no forced order" the *missed-dependency rate*. The 236 questions (132 free, 104 forced; majority baseline 55.9%) split by ground truth into two errors of different nature (Figure 3d; per-model values in Appendix Table 4).

The first is declaring a free order forced: on pairs with no geometric precedence, the rate of asserting one (the invented-dependency rate) lies between 20% and 50% for 14 models, up to 46.2%. The second failure mode is extreme majority-class collapse: while six models achieve an invented-dependency rate below 10%, five do so due to severe degenerate bias, predicting "no forced order" for nearly all pairs (92.4–99.2% on free pairs vs. 0.0–14.4% on forced pairs). Only 5 of 23 models beat the majority baseline on both kinds of pair, and only the best model combines an invented-dependency rate below 10% (8.3%) with forced-pair accuracy near 80% (79.8%). Combined accuracy cannot separate the two errors.

4.4 Separating Perception from Reasoning

The perception–reasoning pairing is built into the question sequence, and it is the reason all questions of an assembly are asked in one shared conversation. Questions go from 2D to 4D within the conversation, and every L2 reasoning question about a part—4D-a mobility (including unique free

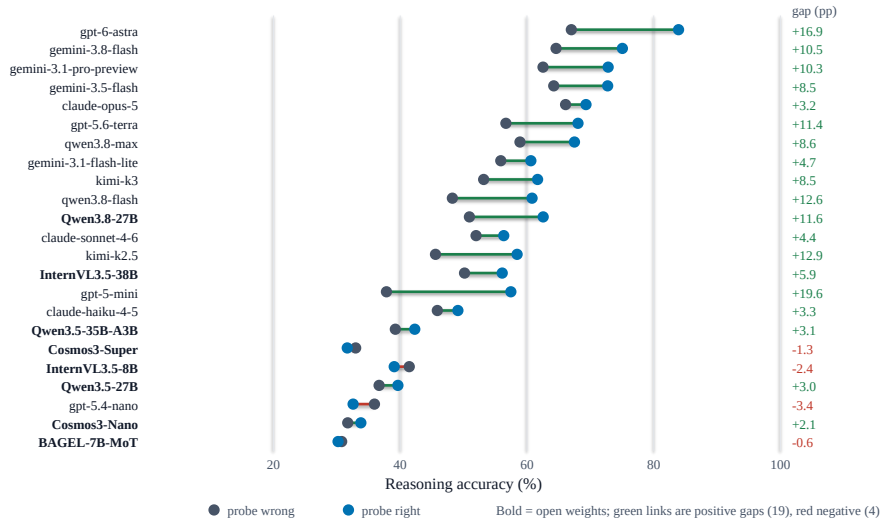

Figure 4: Reasoning accuracy per model after a wrong (gray) and a correct (blue) perception probe on the same part; the link is the gap, green if positive (19 models) and red if negative (4). Bold names are open-weight models.

direction), 4D-b blocker or 4D-c precedence—is preceded by a probe on the same part: a 3D-c contact-count question ("how many parts does Part_k touch?") that probes the model's reading of the local structure around that part. When the reasoning question arrives, the model's earlier probe response remains in the context when it answers the reasoning question; asking each question in a fresh context would neither allow this nor let us relate each reasoning answer to the model's own answer on the matched probe. The 23 models yield 33,313 answered pairs, and in every pair the probe was answered first.

Grouping reasoning answers by the correctness of their probe gives conditional accuracies (Figure 4). Pooled, reasoning accuracy is 57.8% after a correct probe and 47.0% after a wrong one, and the gap holds for every reasoning type: mobility 63.9% vs. 53.7%, unique free direction 50.4% vs. 36.9%, blocker 53.8% vs. 42.2%, precedence 60.4% vs. 54.2%. Per model the gap is positive for 19 of 23 models but limited, 6.7 points on average and 19.6 at most (gpt-5-mini); even the best model reaches only 83.9% after a correct probe (67.0% after a wrong one). The 2×2 split shows how loosely they are coupled: 30.3% of pairs are right on both and 25.2% wrong on both, and the two mixed cells are almost equal—22.1% a correct probe with wrong reasoning, 22.4% the reverse. Reasoning errors therefore persist even when the matched contact-count probes are answered correctly, showing that success on these local structural probes does not ensure correct operation reasoning. Because a correct count does not establish that the geometry relevant to an operation was read correctly (Appendix O), the pairing does not by itself separate misperception from misreasoning.

4.5 Capability Dimension and Difficulty

Accuracy can be read along the capability dimension (2D, 3D, 4D) and along two difficulty axes, the scope of a question (L1–L3, Section 3.1) and the size of the assembly (Figure 5).

Dimension. Pooled accuracy is 55.4% on 2D, 54.0% on 3D and 52.9% on 4D, and only 6 of 23 models are weakest on 4D: the three capability dimensions organize what is asked but do not rank difficulty, in line with Section 4.2.

Scope. Scope separates questions far more. L1 questions, answerable from the named part or pair alone, score 60.4%; L2 questions, which relate a named part to the rest of the structure, 51.0%; L3 questions on the whole structure 55.7%. L1 is the easiest scope for 14 models and L2 the hardest for 13. The pattern repeats within a dimension (Figure 5a): in 3D, pairwise contact (L1) scores 63.8%, contact counting (L2) 49.0% and the most-connected part (L3) 59.6%; in 4D, the L2 types score 52.7% and the L3 types 54.1%. On average, locating one named part within the structure and aggregating its relations (L2) is harder than both reading a named part or pair (L1) and answering

about the whole structure (L3). Since each question type has a single scope, these are differences between the types that make up each scope (Appendix H).

Part Count. Accuracy falls from 59.2% at 4–5 parts to 53.9% at 6–7 and 48.8% at 8–12; 22 of 23 models decline, by up to 24.4 points (gpt-5-mini). The decline is nearly the same everywhere (Figure 5b,c): from 4–5 to 10–12 parts, 10.3–10.7 points in each capability dimension and 7.5–10.1 in each scope, about 2.1–2.4 points per part. Part count thus adds a near-uniform cost on top of scope, including on color questions that need no reasoning, consistent with visual crowding (Campbell et al., 2024); the one exception is pairwise precedence (+0.4 per part, $p = 0.31$), which asks about two parts regardless of the total. Disassembly layers are confounded with part count (Appendix D).

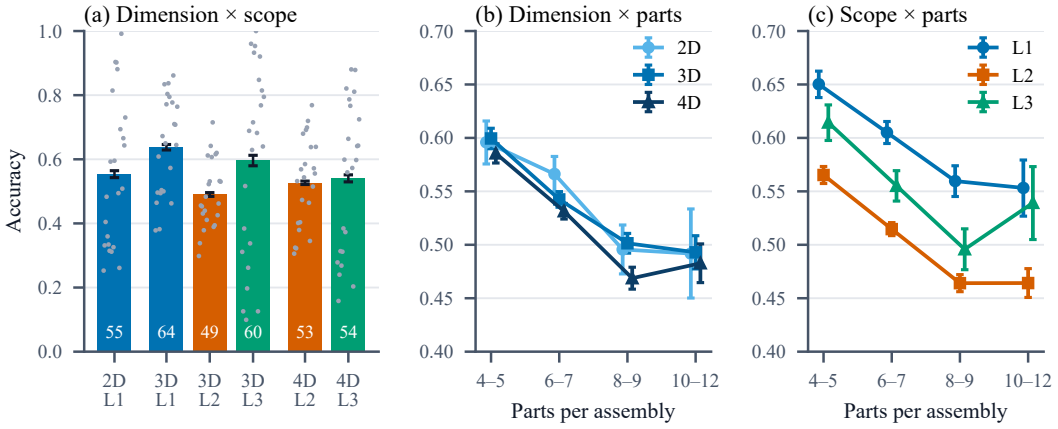

Figure 5: **Accuracy by capability dimension and difficulty**. 23 models pooled, Wilson 95% intervals. (a) Capability dimension × scope; gray dots are single models. (b, c) By part count, per dimension group (b) and per scope (c).

Part Volume. A further difficulty factor is the size of the target, the part named in the question: in 29.9% of questions it is below 2% of the assembly (pins and keys). Small targets are harder on mobility (40.1% vs. 54.6% when removable, 75.8% vs. 81.6% when blocked) and on blocker attribution (40.1% vs. 51.8%), for 18 of 23 models and most of all for the best one, gpt-6-astra (22.4 points), consistent with the size sensitivity of MLLM perception (Zhang et al., 2025a) (Appendix E).

5 Conclusion and Limitations

We presented TenonBench, which measures whether VLMs can infer motion feasibility and operation dependencies from visual observations of interlocking parts, using geometric ground truth, partial-order evaluation and paired perception–reasoning diagnosis. The evaluation of 23 models supports four conclusions. (1) The benchmark is far from solved (best 77.2%, mean 53.7%), and the gap lies in aggregate judgments: naming the part that blocks a direction is on average harder than judging that the direction is blocked, and full-sequence validity (4D-d) is the weakest dimension on average. (2) Pairwise precedence exposes two errors that combined accuracy hides: strong models invent dependencies between free parts, weak models collapse onto "no forced order", and only one model avoids both. (3) Reasoning accuracy is higher when the matched contact-count probes are correct, but only by 6.7 points on average, and reasoning errors persist even when these probes are answered correctly: success on local structural probes does not ensure correct operation reasoning. (4) Difficulty follows scope and part count rather than the 2D/3D/4D grouping: questions that relate one part to the rest of the structure are the hardest, each additional part costs about two points in every dimension, and small parts remain a weakness even for the best model. Because the assets are parametric 3D models, future versions can add multi-view and interactive settings in simulation and give the solver richer motion primitives. It also measures what a model recognizes rather than what it executes; connecting its diagnoses to closed-loop manipulation, for which the simulated robot-arm reassembly (Appendix L) is a starting point, would show whether these gaps predict the failures of embodied agents. Beyond evaluation, the solver's blocking labels and partial orders could serve as supervision for teaching models the constraints they now invent or miss.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing and proofreading, writing data-analysis and plotting scripts, and searching the literature. We did not use generative AI tools to produce the benchmark questions or their ground truth: all labels are computed by the geometric solver and all questions are instantiated from deterministic templates (Section 3.2); the 23 VLMs appear only as evaluated subjects. We have reviewed all AI-assisted work: every reported number was recomputed from the stored model answers by the analysis scripts, and references added with AI assistance were checked against their publisher or proceedings pages. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The data are procedurally generated mortise-and-tenon 3D models and involve no human subjects or personal information.

REPRODUCIBILITY STATEMENT

Question generation calls no language model and its seed is derived from the assembly name, so the bank is fully reproducible from the 3D assets. Scoring is decoupled from model calls. The precedence solver and the feasibility replay are independent implementations checked for agreement. The appendix gives protocol details, solver thresholds, examples per type, per-model tables and the raw answers of the failure case.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 846–858. IEEE, 2021.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*, 2025.

Aditya Chetan, Eric Cai, Peeyush Kushwaha, Bharath Raj Nagoor Kani, Utkarsh Mall, Qianqian Wang, Noah Snavely, and Bharath Hariharan. Flat-Pack Bench: Evaluating spatio-temporal understanding in large vision-language models through furniture assembly. *arXiv preprint arXiv:2605.21625*, 2026.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, volume 2025, pp. 97959–98108, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. Every mistake counts in assembly. *arXiv preprint arXiv:2307.16453*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. *arXiv preprint arXiv:2502.04313*, 2025.

Sanjay Haresh, Daniel Dijkman, Apratim Bhattacharyya, and Roland Memisevic. Clevrskills: Compositional language and visual reasoning in robotics. *Advances in Neural Information Processing Systems*, 37:38235–38266, 2024.

Haochen Huang, Jiahuan Pei, Mohammad Aliannejadi, Xin Sun, Moonisa Ahsan, Chuang Yu, Zhaochun Ren, Pablo Cesar, and Junxiao Wang. Lego co-builder: exploring fine-grained vision-language modeling for multimodal lego assembly assistants. *arXiv preprint arXiv:2507.05515*, 2025.

Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 135833–135865, 2026.

Zhi Jing, Jinbin Qiao, Ouyang Lu, Jicong Ao, Shuang Qiu, Huazhe Xu, Yu-Gang Jiang, and Chenjia Bai. Assemlm: A spatial reasoning multimodal large language model for robotic assembly. *arXiv preprint arXiv:2604.08983*, 2026.

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

Tina Khezresmaeilzadeh, Jike Zhong, and Konstantinos Psounis. VRIQ: Benchmarking and analyzing visual-reasoning IQ of VLMs. *arXiv preprint arXiv:2602.05382*, 2026.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Hao Li, Zi-han Ding, Shuai Yang, Yilun Chen, Yang Tian, Xiaolin Hu, Tai Wang, Dahua Lin, Feng Zhao, Si Liu, et al. RoboInter: A holistic intermediate representation suite towards robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 113314–113381, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Mingfei Han, et al. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly. *Advances in Neural Information Processing Systems*, 38, 2026a.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. CausalSpatial: A benchmark for object-centric causal spatial reasoning. *arXiv preprint arXiv:2601.13304*, 2026b.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. iVISPAR – an interactive visual-spatial reasoning benchmark for VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26745–26769, 2025.

Cheolhong Min, Jaeyun Jung, Daeun Lee, Hyeonseong Jeon, Yu Su, Jonathan Tremblay, Chan Hee Song, and Jaesik Park. Why far looks up: Probing spatial representation in vision-language models. In *European Conference on Computer Vision*, pp. 22–39. Springer, 2026.

Nhi Pham and Michael Schott. H-POPE: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. *arXiv preprint arXiv:2411.04077*, 2024.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of VLMs. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2024.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision*, pp. 293–309. Springer, 2024.

Silvia Sellán, Yun-Chun Chen, Ziyi Wu, Animesh Garg, and Alec Jacobson. Breaking bad: A dataset for geometric fracture and reassembly. *Advances in Neural Information Processing Systems*, 35: 38885–38898, 2022.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21064–21074. IEEE, 2022.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning?(2025). *arXiv preprint arXiv:2503.19990*, 2025.

Chenrui Tie, Shengxiang Sun, Jinxuan Zhu, Yiwei Liu, Jingxiang Guo, Yue Hu, Haonan Chen, Junting Chen, Ruihai Wu, and Lin Shao. Manual2Skill: Learning to read manuals and acquire robotic skills for furniture assembly using vision-language models. *arXiv preprint arXiv:2502.10090*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.

Qineng Wang, Wenlong Huang, Yu Zhou, Hang Yin, Tianwei Bao, Jianwen Lyu, Weiyu Liu, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, et al. ENACT: Evaluating embodied cognition with world modeling of egocentric interaction. In *International Conference on Learning Representations*, volume 2026, pp. 153022–153077, 2026.

Ruocheng Wang, Yunzhi Zhang, Jiayuan Mao, Ran Zhang, Chin-Yi Cheng, and Jiajun Wu. IKEA-Manual: Seeing shape assembly step by step. *Advances in Neural Information Processing Systems*, 35:28428–28440, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, Mahfuz Ahmed Anik, Munem Shahriar, Mohsin Mahmud Topu, Sadia Tasnim Meem, Rahatun Nesa Priti, Sabrina Afroz Mitu, Md Iqramul Hoque, et al. SpatiaLab: Can vision-language models perform spatial reasoning in the wild? *arXiv preprint arXiv:2602.03916*, 2026.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025a.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025b.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

A Protocol Details

Evaluation is per assembly: all questions of one assembly are asked in one shared conversation, 178 conversations in total, 18–30 questions each (median 23), one question per turn. The images—one line-art whole-assembly render, one whole-assembly render in the fixed palette, and the per-part views of all parts—are given once in the first turn; later turns contain only the question text, and the context keeps all earlier exchanges. Questions are ordered by increasing dimension, and the paired perception probe always precedes its reasoning question (all 782 pairs). Besides the option, the model must first give a one-sentence rationale grounded in geometric evidence from the images, which enables post-hoc inspection of failure patterns. Scoring compares the option with the programmatic label; if a question has several equivalent correct formulations all are accepted; failed calls and unparsable outputs are excluded from the denominator; choosing "unsure / cannot be determined from the images" counts as wrong and is also reported as abstention. Prompts are in English to avoid confounding with language-specific instruction following. The reported numbers are therefore obtained under shared context within an assembly in one continuous shared conversation. Within an assembly, perception questions precede reasoning questions within a dimension and dimensions go 2D, 3D, 4D. The effect of within-conversation position on accuracy is confounded with this difficulty order, and the residual trends within a question type have mixed signs, so we draw no conclusion.

B Question Counts and Results per Type

For the ten question types of Section 3.1, Table 3 lists counts, random baselines, the mean of 23 models and the best model. The seven sub-dimensions of Table 2 are aggregations of these ten types.

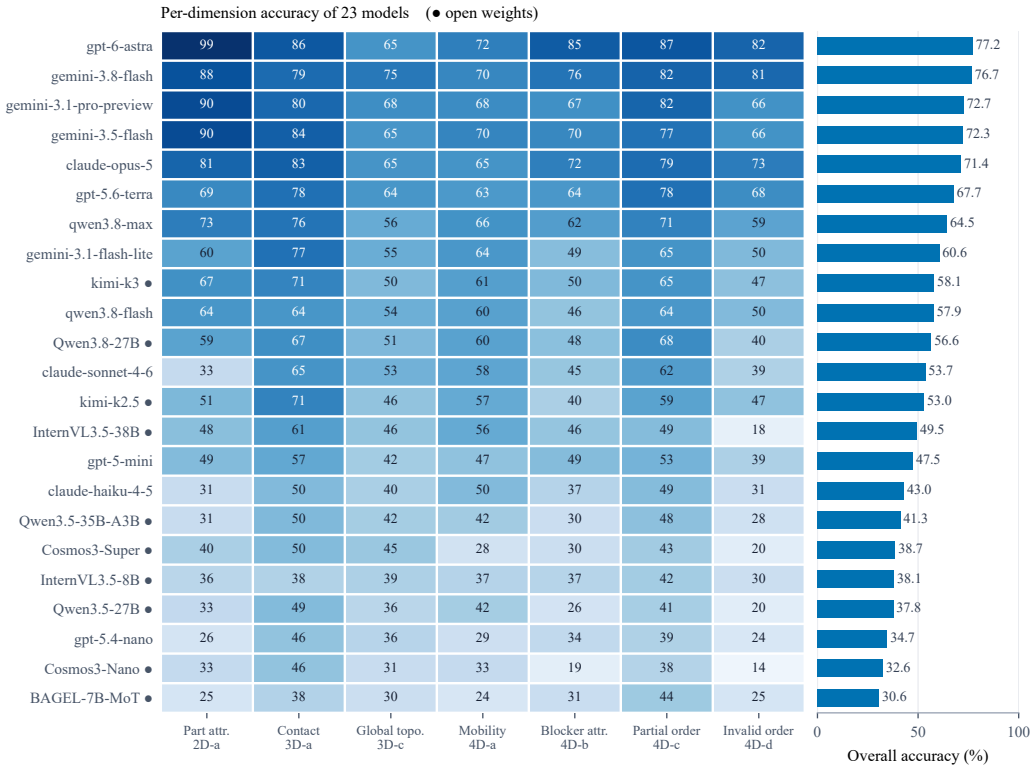

Figure 6: Accuracy (%) of the 23 models on the seven sub-dimensions, same source as Table 2, ordered by overall accuracy; the bars on the right are overall accuracy and the dot after a name marks open-weight models. The lightest cell of almost every row falls in 3D-c, 4D-b or 4D-d.

Scale and generation within families. The three open families agree (main-text Figure 3): InternVL3.5 38B 49.5% versus 8B 38.1%; Cosmos3 Super 38.7% versus Nano 32.6%; Qwen goes from 37.8% (3.5-27B) to 41.3% (3.5-35B-A3B) to 56.6% (3.8-27B), so generational gains exceed within-generation size gains. 4D-d is the lowest dimension of all seven models; across Qwen generations it rises from 19.5% to 40.1%, but scaling within a generation need not help: InternVL3.5-38B scores 17.5% on 4D-d, below the 29.9% of its 8B version. The best open model, Qwen3.8-27B, ranks 11th at 56.6%, more than 15.7 points behind the top four closed models. Across models, mean output length correlates weakly with overall accuracy (Spearman $\rho = 0.43$, $p = 0.04$): the overall best model averages 47 tokens, while the longest-output model (2,106 tokens on average) scores 4.5 points lower; within-model output lengths of correct versus wrong answers are in Appendix H. Model rankings agree across dimensions: the median Kendall τ between the rankings of any two dimensions is 0.77 (minimum 0.69 between 2D-a and 4D-d, maximum 0.85 between 3D-c and 4D-c); gpt-6-astra leads six of the seven dimensions and gemini-3.8-flash leads 3D-c (74.7%).

Table 3: Results per question type (%). Mean is the average accuracy of 23 models; Best is the best model. Counts are unbalanced, so no cross-type total is averaged; color questions are limited by the coverage of the colored render.

Question type	Dim.	#Q	Random	Mean	Best
Color of a named part	2D-a	356	20.3	55.6	99.2
Whether two parts touch	3D-a	534	33.3	63.9	86.1
Contact count of a named part	3D-c	1,200	20.0	49.1	71.5
Most-connected part	3D-c	151	20.0	59.6	100.0
Removable along a given direction	4D-a	534	33.3	59.1	76.6
Unique free direction	4D-a	344	12.5	43.9	68.0
Blocking part	4D-b	356	20.0	48.4	84.8
First-removable part	4D-c	177	24.4	63.7	95.5
Forced pairwise order	4D-c	236	25.0	57.5	81.4
Infeasible order among four	4D-d	177	20.0	44.2	81.9

C Per-Model Results on Pairwise Precedence

Table 4 gives the per-model values behind Section 4.3: accuracy on free and on forced pairs, together with the invented-order and missed-dependency rates.

D Details on Scale and Constraint Strength

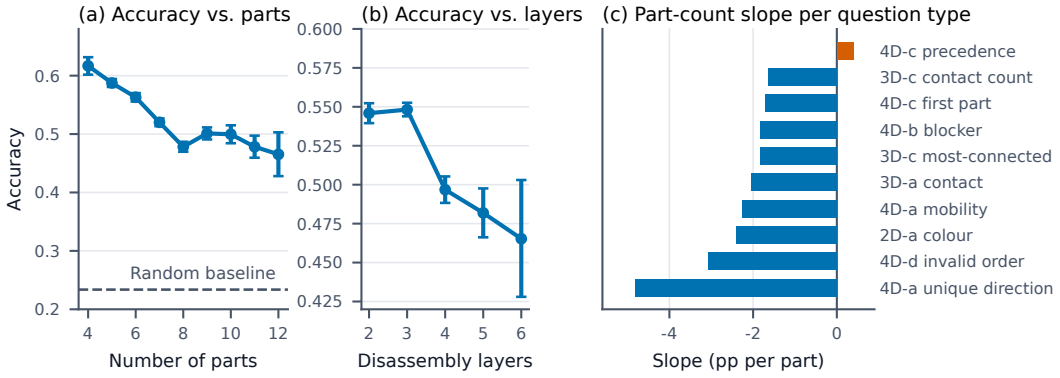

Figure 7: Details of the scale effect in Section 4.5; 23 models pooled, Wilson 95% intervals. (a) Accuracy by part count; dashed line is the random baseline including abstention. (b) Accuracy by disassembly layers (1 layer has only 432 answers and is omitted). (c) Linear slope of accuracy against part count within each question type (points per part); pairwise precedence is the only non-negative one (orange).

Table 4: Pairwise precedence split by ground truth (%), sorted by forced-pair accuracy. 132 free pairs and 104 forced pairs; majority baseline 55.9%. Invented-dependency rate: asserting an order on a free pair. Missed-dependency rate: answering "no forced order" on a forced pair. The upper block exceeds the majority baseline on forced pairs, the lower block is biased; a severe imbalance between the two accuracy columns is the degenerate one-sided behavior. Bold marks the best value per column within the upper block (highest for the first two, lowest for the last two); the high free-pair accuracy of the lower block comes from always answering "no forced order" and is not comparable.

Model	Free acc.	Forced acc.	Invented	Missed
gemini-3.5-flash	51.5	**92.3**	46.2	**3.8**
gemini-3.8-flash	62.9	84.6	34.1	5.8
claude-opus-5	60.6	82.7	39.4	17.3
gemini-3.1-pro-preview	73.4	80.4	21.1	9.8
gpt-6-astra	**82.6**	79.8	**8.3**	19.2
gpt-5.6-terra	56.1	76.9	43.9	23.1
gemini-3.1-flash-lite	56.8	47.1	43.2	51.0
qwen3.8-max	73.5	44.2	24.2	52.9
claude-sonnet-4-6	59.8	44.2	37.1	44.2
gpt-5-mini	41.7	39.8	29.5	38.8
kimi-k2.5	59.1	37.5	37.9	58.7
kimi-k3	72.7	29.8	15.2	66.3
qwen3.8-flash	74.2	26.9	23.5	72.1
claude-haiku-4-5	75.0	25.0	25.0	69.2
Qwen3.8-27B	95.5	14.4	4.5	83.7
Qwen3.5-35B-A3B	74.0	14.4	26.0	76.0
Qwen3.5-27B	63.6	11.5	31.8	76.9
InternVL3.5-38B	85.6	9.6	14.4	86.5
Cosmos3-Super	87.9	9.6	12.1	84.6
BAGEL-7B-MoT	92.4	3.8	7.6	94.2
InternVL3.5-8B	98.5	1.9	1.5	96.2
Cosmos3-Nano	92.4	1.9	7.6	96.2
gpt-5.4-nano	99.2	0.0	0.8	98.1

Per model and per type. Drop between the 4–5 and $\geq$8 part bins: largest for gpt-5-mini, 24.4 points (59.3% to 34.9%); smallest for gpt-5.4-nano, -3.7 points (33.0% to 36.7%), a model near random in both bins. Within-type slopes per part: unique free direction -4.8, invalid sequence -3.1, color -2.4, mobility -2.3, contact pair -2.0, most connected -1.8, blocker -1.8, first removable -1.7, contact count -1.6, all $p < 10^{-4}$; pairwise precedence $+0.4$ ($p = 0.31$). The plateau at 9–11 parts reflects counts (396, 179, 117 questions); 12 parts has only 30.

Disassembly layers and number of valid orders. Pooled accuracy for 2 to 6 disassembly layers is 54.6%, 54.8%, 49.7%, 48.2%, 46.5%, with abstention 4.5%, 4.5%, 6.6%, 10.3%, 13.2%, and a slope of -2.2 points per layer. Disassembly depth correlates with part count and are not independent evidence. L3 questions binned by number of valid orders: $\leq$10 orders 54.0%, 10–100 61.5%, 1k–10k 49.2%, $>$10k 54.1%, not monotone; the slope against $\log_{10}$ of the count is -2.5 points, with invalid sequence identification at -4.0 (more valid orders make the single infeasible one harder to spot) and pairwise precedence at $+1.5$.

Full-sequence validity: later violations are harder to find. When the infeasible step of the invalid sequence is step 1, accuracy is 55.6%; step 2, 49.8%; step 8, 33.3%; step 9, 25.0%; -3.0 points per step. When the violating part sits in disassembly layer 1, 2 or 3, accuracy is 40.6%, 46.5%, 33.0%. Models tend to check only the first few steps. This variable nearly coincides with part count (-3.1 per part) and cannot be separated from it.

First-removable part: more acceptable answers do not make it easier. The task has 4.49 acceptable answers on average. With one acceptable answer accuracy is 76.5%, with six 58.3%: structures with many acceptable answers have more parts and are looser, offsetting the easier guess.

E Details on Volume and Visibility

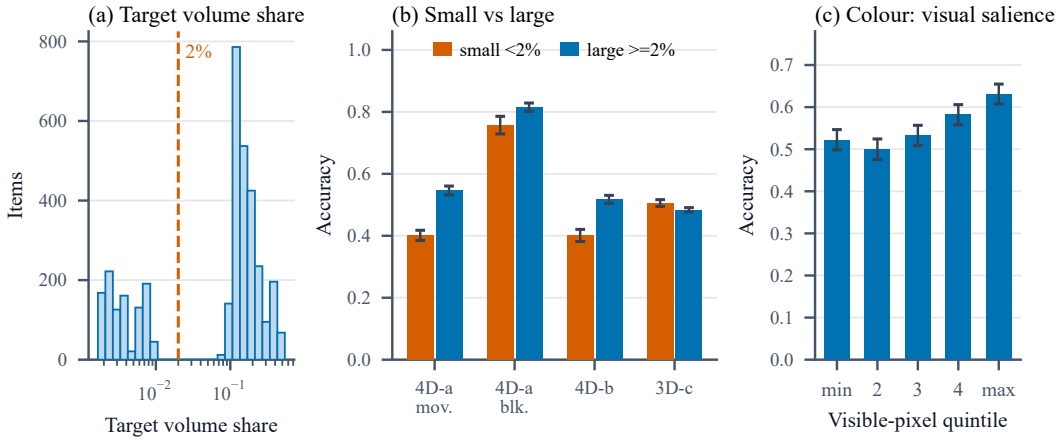

Figure 8: (a) Distribution of target-part volume share over the 3,560 questions with a target part (log axis); dashed line at 2%. (b) Small targets (volume share below 2%) versus large ones; mobility questions are split by ground truth. (c) Color questions by quintile of the target's visible-pixel share in the colored render; Wilson 95% intervals.

Distribution and per model. The deciles of volume share are 0.17%, 0.25%, 0.56%, 7.9%, 11.1%, 12.5%, 13.9%, 16.5%, 20.0%, 27.0%; almost no question lies between 2% and 10%, so the definition of "small part" does not depend on the exact threshold; 1,065 questions have a small target. 18 of 23 models do worse on small parts; the four largest gaps are gpt-6-astra 22.4, gpt-5-mini 18.3, qwen3.8-flash 18.0 and gemini-3.1-pro-preview 17.3 points; the five models with zero or reversed gaps (Cosmos3-Nano -4.2, gpt-5.4-nano -3.6, InternVL3.5-8B -2.5, claude-haiku-4-5 -0.2, Qwen3.5-27B -0.2) all have overall accuracy at most 43%. After binning by part count, 8 of 9 bins favor large parts (the exception is 4 parts, with only 138 small-part answers).

Visibility. Color questions target visible parts only and contain no small parts; by quintile of visible-pixel share, accuracy is 52.3%, 50.0%, 53.3%, 58.2%, 63.1%, and 21 of 23 models are better on the top quintile than on the bottom one. This is the difficulty of "visible but small", not occlusion.

Two types that are not volume effects. Contact counting by volume quartile gives 50.0%, 54.2%, 52.5%, 39.4%, worst for the largest quartile: large parts touch more parts, and accuracy is 27.6–39.7% when the label is 3 or more against 53.5–57.4% when it is 1 or 2, the same phenomenon as the shrinkage in Appendix F. On unique-free-direction questions small parts score 21.4% and large 53.4%, but 84.6% of small-part labels are $\pm Y$ and none is $+Z$, whereas 74.6% of large-part labels are $+Z$; with the label fixed to $+Z$, the four volume quartiles score 62.6%, 68.0%, 67.1%, 64.6% and the volume effect disappears. The small-part deficit on this type is another appearance of the directional prior of Appendix F, so the main text does not use it as evidence for a volume effect.

Two shortcut checks related to volume. In 54.5% of blocker questions the true blocker is the largest part; these score 51.8% and the rest 44.2%. But when the blocker is not the largest part, models pick the largest only 13.9% of the time, so the gap comes from small blockers being hard to identify, not from a "big blocks small" heuristic. In 96.0% of the 151 most-connected questions the label is the largest part; on the 6 questions where they differ, models pick the largest only 8.7% of the time—the shortcut exists but is not used. The correlation between label and volume in this type is a defect of this version of the bank; the next version should add instances where volume and contact count disagree.

F SYSTEMATIC PRIORS AND BIASES

Errors concentrate on few options (the most common wrong option takes 65.4% of a question's errors, Appendix H), and two priors are measurable from the existing answers.

The first is "pull it up". A unique-free-direction question (a 4D-a variant) states that a part can be removed along exactly one direction and asks which of $\pm X$, $\pm Y$, $\pm Z$ it is. Accuracy is 65.6% when the ground truth is $+Z$ and 19.6% over the other five directions (Figure 9a); 20 of 23 models are more accurate on $+Z$, by up to 76.7 points; $+Z$ makes up 41.9% of all answers and $-Z$ only 3.0%. The prior matches the assets—tenons are mostly inserted from above and $+Z$ has the lowest blocked rate (Appendix J)—and models learned the statistic (Geirhos et al., 2020; Min et al., 2026) rather than checking blockers: with the label fixed to $+Z$, accuracy no longer varies with target volume (Appendix E).

The second is under-answering the positive side on the two yes/no tasks, contact (3D-a: do two parts touch?) and mobility (4D-a: can a part be removed along a direction?): 66.7% of contact labels are "in contact" but models say so 47.6% of the time; 66.6% of mobility labels are "removable" but models say so 38.3% of the time; both hold for all 23 models, the opposite of the 'yes' bias on object existence (Li et al., 2023) and in line with the 'no' bias on attributes (Pham & Schott, 2024). A part with fewer contacts should be more mobile, so under-estimating both is not blanket conservatism: the two judgments a model makes about the same parts are inconsistent. Contact-count answers likewise shrink towards 1–2 (Appendix F).

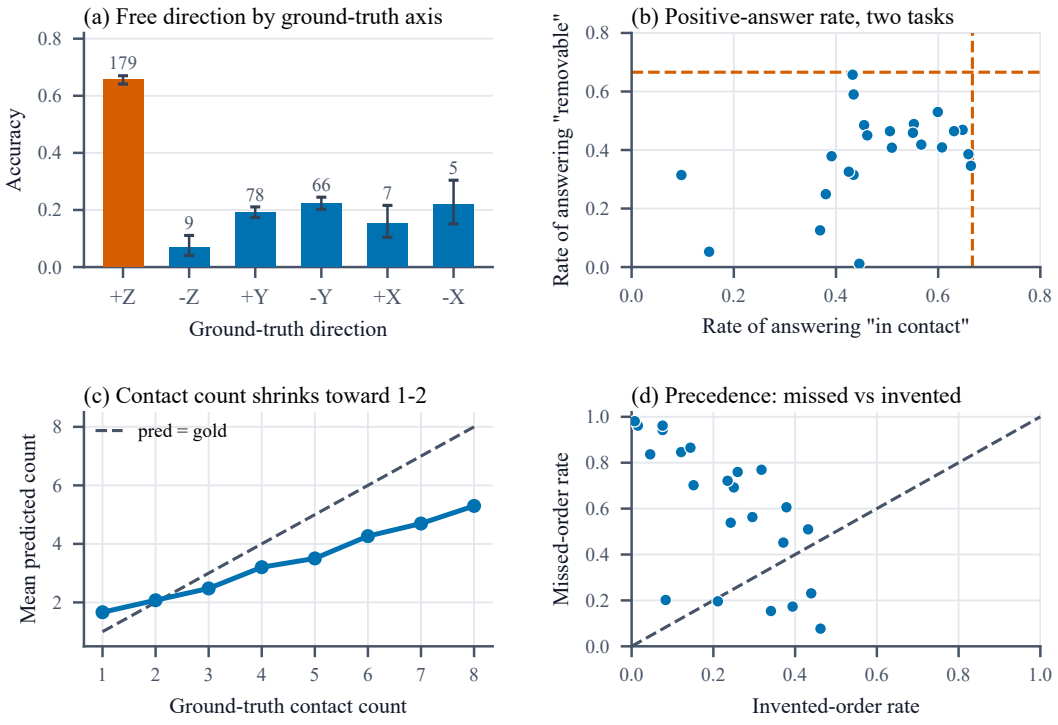

Figure 9: (a) Unique-free-direction accuracy by ground-truth direction, question counts above the bars, $+Z$ in orange. (b) One point per model: rate of answering "in contact" against rate of answering "removable"; orange dashed lines are the base rates (66.7% and 66.6%); all 23 points lie in the lower left. (c) Mean predicted contact count by ground truth; dashed line is prediction equals truth. (d) One point per model: invented-dependency rate against missed-dependency rate on pairwise precedence; dashed line is the diagonal.

Pull it up. Question counts for ground truth $+Z$, $-Z$, $+Y$, $-Y$, $+X$, $-X$ are 179, 9, 78, 66, 7, 5, with accuracy 65.6%, 6.8%, 19.2%, 22.3%, 15.2%, 21.8%. The last three have few questions and wide intervals; the conclusion rests on $+Z$ versus non-$+Z$. Per model, the $+Z$ minus non-$+Z$

Table 5: Systematic effects per model (%). Small-part gap is large-part minus small-part accuracy (positive: small parts harder); $+Z$ gap is accuracy on $+Z$ minus non-$+Z$ labels on unique-free-direction questions; invented and missed rates as in Table 4. Sorted by overall accuracy within each block.

Model	Overall	Small-part gap	$+Z$ gap	Invented	Missed	Abstention
Closed API						
gpt-6-astra	77.2	+22.4	+64.3	8.3	20.2	8.4
gemini-3.8-flash	76.7	+13.6	+66.5	34.1	15.4	6.3
gemini-3.1-pro-preview	72.7	+17.3	+71.0	21.1	19.6	8.5
gemini-3.5-flash	72.3	+7.6	+63.2	46.2	7.7	6.5
claude-opus-5	71.4	+7.5	+76.4	39.4	17.3	0.6
gpt-5.6-terra	67.8	+4.5	+52.4	43.9	23.1	0.8
qwen3.8-max	64.5	+16.5	+61.4	24.2	53.8	7.6
gemini-3.1-flash-lite	60.6	+8.9	+76.7	43.2	51.0	3.4
kimi-k3	58.1	+14.1	+63.0	15.2	70.2	7.9
qwen3.8-flash	57.9	+18.0	+62.0	23.5	72.1	2.3
claude-sonnet-4-6	53.7	+2.3	+64.5	37.1	45.2	3.4
kimi-k2.5	53.0	+11.4	+49.0	37.9	60.6	8.4
gpt-5-mini	47.5	+18.3	+67.3	29.5	56.3	26.7
claude-haiku-4-5	43.0	-0.2	+42.9	25.0	69.2	2.8
gpt-5.4-nano	34.7	-3.6	-0.5	0.8	98.1	0.6
Open weights						
Qwen3.8-27B	56.6	+16.0	+65.7	4.5	83.7	4.9
InternVL3.5-38B	49.5	+2.9	+29.5	14.4	86.5	0.9
Qwen3.5-35B-A3B	41.3	+0.7	+38.9	25.9	76.0	1.6
Cosmos3-Super	38.7	+3.6	+3.1	12.1	84.6	2.9
InternVL3.5-8B	38.1	-2.5	+5.7	1.5	96.2	0.2
Qwen3.5-27B	37.8	-0.2	+42.3	31.8	76.9	6.8
Cosmos3-Nano	32.6	-4.2	-2.7	7.6	96.2	4.2
BAGEL-7B-MoT	30.6	+4.4	-7.0	7.6	94.2	1.9

gap is positive for 20 models, the largest four being gemini-3.1-flash-lite 76.7, claude-opus-5 76.4, gemini-3.1-pro-preview 71.0 and gpt-5-mini 67.4 points; the three reversed models (gpt-5.4-nano -0.5, Cosmos3-Nano -2.7, BAGEL-7B-MoT -7.0) score at most 22% on this type. $+Z$ is the label of 52.0% of these questions, so "always $+Z$" is a 52.0% majority baseline, above the type's pooled accuracy of 43.8%.

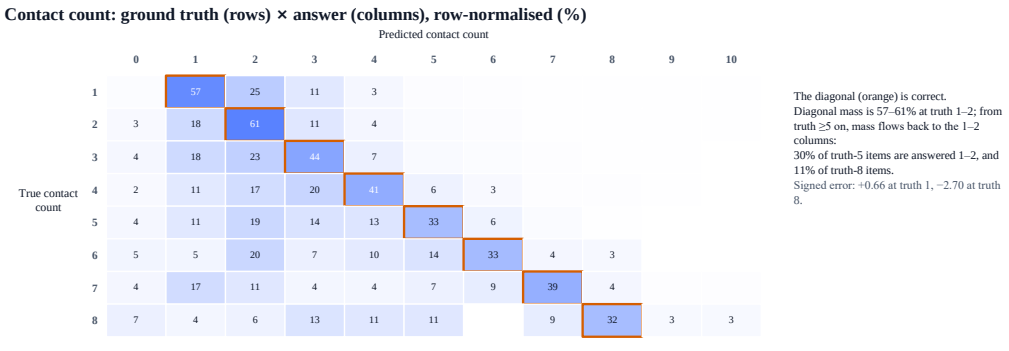

Figure 10: Confusion matrix of contact counting: ground truth (rows) against answer (columns), row-normalized to percent, orange frames on the diagonal. Diagonal mass is 57–61% at truth 1 or 2; from truth 5 on, answers flow back to the 1–2 columns.

Under-answering the positive side and contact-count shrinkage. Contact questions score 59.2% when the label is "in contact" and 72.9% when it is not; mobility questions score 48.5% when the label is "removable" and 80.3% when it is "blocked". Signed error of contact counts by ground truth (Figure 10): $+0.66$ at 1, $+0.07$ at 2, -0.52 at 3, -0.80 at 4, -1.50 at 5, -2.70 at

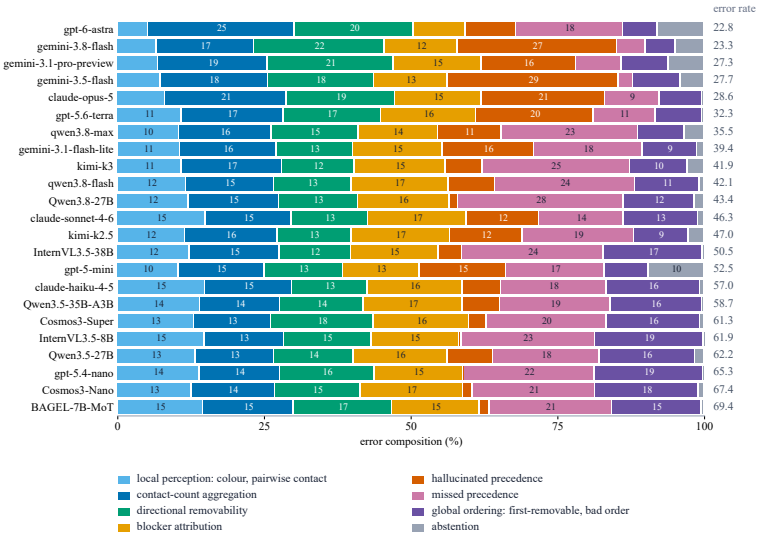

Figure 11: Composition of errors of the 23 models over the eight root causes of Appendix G, ordered by total error rate (right end, %).

8; 64.7% of answers are 1 or 2. Per-model small-part gaps, $+Z$ gaps, invented and missed rates and abstention are listed in Table 5. The rise of mobility accuracy with the target's contact count (59.6% at 1 contact, 83.3% at 4) is not a capability signal: parts with 4 or more contacts are almost all labeled "blocked" (the item-level removable share falls from 78.1% at 1 contact to 6.7% at 4), in the same direction as the models' bias toward "blocked". By ground-truth axis, mobility accuracy is 70.1% on X, 63.1% on Z and 48.6% on Y, without polarity control here.

Precedence: more missed than invented. Pooled numbers for the two errors of Section 4.3: free-pair accuracy 73.5%, forced-pair accuracy 37.3%; missed-dependency rate 59.8%, invented-dependency rate 23.3%; only 5 models invent more than they miss; wrong direction (an order asserted but reversed) is only 7.3% of forced-pair answers. Per-model values are in Table 4.

G Error Spectrum by Root Cause

The error spectrum assigns every wrong answer to the ability its type tests: local perception (color, pairwise contact), contact aggregation, directional mobility, blocker attribution, whole-structure questions (most connected, first removable, invalid sequence) and abstention, with pairwise precedence split into invented and missed dependencies as defined in Section 4.3. Error rates are computed per cause and normalized to 100% per model (Figure 11). The two precedence errors move in opposite directions: for the six models with the lowest total error, invented dependencies make up 20.3% of the composition and missed dependencies 8.9%; for the last seven it is 3.4% and 20.3%. gpt-6-astra is the exception (18.3% missed against 8.7% invented). Strong models fail by declaring constraints where none exist, weak models by not seeing them. The global-sequence segment grows monotonically with weakness, from 5.0% at the top to 18.5% at the bottom; blocker attribution stays within 8.7–17.5% for all 23 models and does not improve with overall score. A verbatim example of an invented dependency is in Appendix O: four models assert contradictory orders on the same free pair, and three invent the same non-existent contact—the error begins in grounding, not reasoning (Li et al., 2023; Tong et al., 2024).

H Item Quality and Robustness

Distribution of item difficulty. Of 4,065 questions, 13 (0.3%) are answered by no model and 16 (0.4%) by all 23; the mode lies at 13–15 models. Mean item accuracy is 53.6% with a standard deviation of 20.5 points; 5.9% of items are below 20% and 9.5% above 80%. Of the 13 unsolved

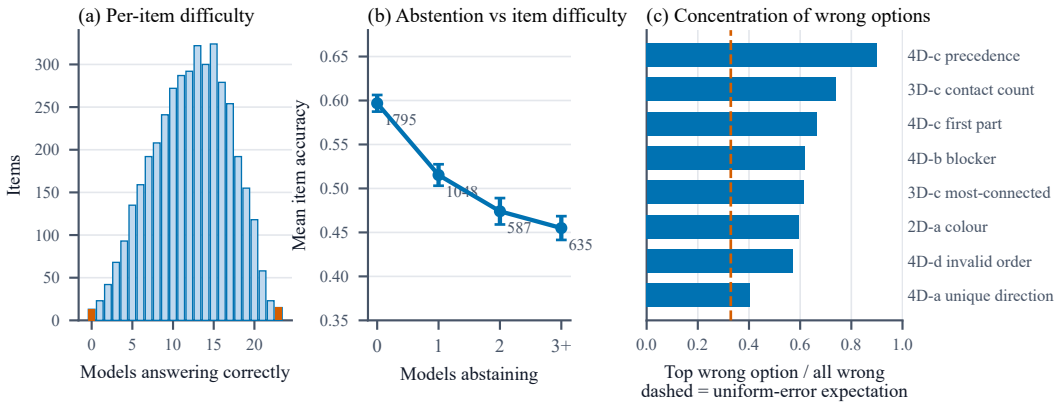

Figure 12: (a) Number of models answering each question correctly: 13 questions by none and 16 by all (orange). (b) Mean item accuracy by the number of models abstaining on the item; labels are item counts. (c) Mean share of a question's errors taken by its most common wrong option, per type; dashed line is the expectation under uniform errors; binary types are excluded.

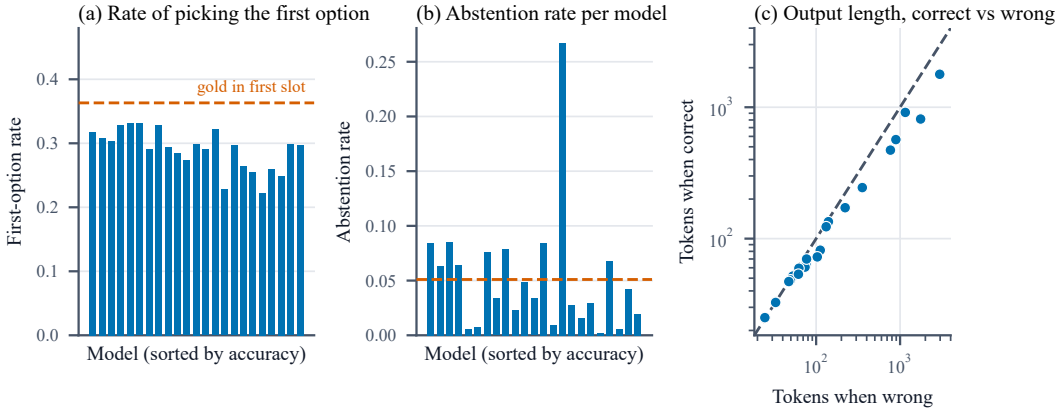

Figure 13: (a) Share of answers on the first option per model; the orange dashed line is the share of labels in the first slot, 36.3%; all 23 models are below it. (b) Abstention per model; dashed line is the pooled rate, 5.1%. (c) One point per model: mean output tokens when wrong against when correct (log–log); dashed line is equality; models below the diagonal write more when wrong.

items, 11 are unique-free-direction questions (3.2% of that type), consistent with the directional prior of Appendix F; the labels of these 13 items are queued for re-inspection in the next version.

Concentration of errors. Over the 2,670 questions with at least one wrong answer, the most common wrong option takes 65.4% of a question's errors on average (32.9% under uniform errors), and in 70.3% of questions more than half of the errors fall on one option. By type: pairwise precedence 90.2% (almost all errors are "no forced order"), contact count 73.9%, first removable 66.4%, blocker 61.7%, most connected 61.3%, color 59.7%, invalid sequence 57.0%, unique free direction 40.4%. Errors are systematic, not noise; converging errors across models are also reported by Goel et al. (2025).

Abstention. Pooled abstention is 5.1%; only gpt-5-mini exceeds 10% (26.7%) and 5 models are below 1%. Items on which no model abstains have a mean accuracy of 59.7%; the 635 items on which three or more models abstain have 45.5% (Spearman -0.29): abstained items are indeed harder, but abstention is too rare overall to serve as a calibration measure.

Option position and output length. The label is in the first slot for 36.3% of questions (positions were not fully shuffled; to be fixed in the next version); models pick the first option 30.6% of the time pooled, and all 23 models are below the label rate, so there is no first-option bias of the kind

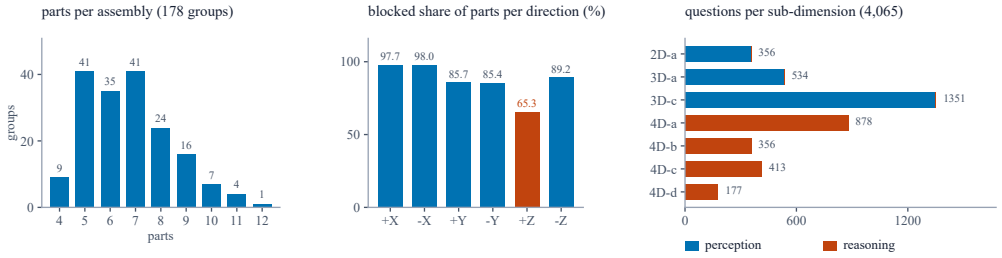

Figure 14: Evaluation set at a glance; all three panels recomputed from the 178 label files and the question bank. Left: parts per assembly. Middle: blocked rate of parts in the six directions; $+Z$ is the only clearly lower direction. Right: question counts per sub-dimension, blue for perception and orange for reasoning.

reported by Zheng et al. (2024). Output length must be read at two levels: pooled, correct answers average 304 tokens and wrong ones 284, but this is a Simpson effect; within a model, wrong answers are longer (Chen et al., 2025), especially for reasoning models (claude-opus-5 472 correct vs. 767 wrong; gemini-3.1-pro-preview 1,783 vs. 2,965; gemini-3.8-flash 813 vs. 1,756). No answer was truncated.

Scope by question type. L1: color 55.4%, pairwise contact 63.8%. L2: contact count 49.0%, mobility 59.1%, unique free direction 43.8%, blocker 48.3%, pairwise precedence 57.5%. L3: most connected 59.6%, first removable 63.7%, invalid sequence 44.0%. Pooled, L1 is 60.4%, L2 51.0% and L3 55.7%. Every type has a single scope, so the scope differences of Section 4.5 are differences between these mixtures of types.

I Design of the Manual Ablation and Prior-Control Conditions

The following two experiments were not run in this version; their designs are fixed here.

Manual ablation. With the question unchanged, provide or withhold step-by-step assembly manual images, compare accuracy on the same questions, and decompose the gain by dimension.

Prior-control conditions. The assets of Section 3.2 have no texture or object semantics; the effect of this on shortcuts is to be tested with two control conditions. The first removes all images and gives only the question and options; the second keeps the images but permutes the mapping between per-part views and the part IDs in the question. If accuracy falls to random under both, models have no directly applicable prior on this benchmark and must solve the questions from image evidence. The $+Z$ prior measured in Appendix F does not depend on images and is expected to survive both conditions; the second condition is additionally re-scored against labels recomputed under the permuted IDs, and the difference between the two scorings is the contribution of grounding IDs to the images.

J Blocking Statistics of the Evaluation Set

Duality of blocking. Each part is tested for translational interference along the six orthogonal directions: if part A moving along d interferes with B, a blocking entry (B, A, d) is recorded, read as B blocks A along d. The relation comes in pairs under direction reversal—A moving along $+d$ into B and B moving along $-d$ into A describe the same interference—so (B, A, d) and $(A, B, -d)$ always hold together. The set has 10,900 blocking entries with a 100% pairing rate, and the counts for $\pm X$, $\pm Y$, $\pm Z$ are exactly equal within each pair: 2,111, 1,549, 1,790 each. This equality is an identity of the relation, not evidence that the assets are unbiased in direction: totals are forced equal by duality, and any directional skew can appear only in how the entries distribute over parts, i.e. the part-level blocked rate, which the next two paragraphs report by axis and within axis.

Table 6: Blocking decisions of the evaluation set (178 assemblies, 1,200 parts) aggregated by axis ($\pm X$ into X, likewise Y and Z). Entries are directional blocking records between part pairs; the counts of the two opposite directions within an axis are identical by duality, so the share column reflects only differences between axes. Part blocked rate is the share of part-level states blocked along either direction of the axis among all $1{,}200 \times 2$ states of that axis; the last column is the difference in blocked rate between the two opposite directions, and Z is the only markedly asymmetric axis.

Axis	Entries	Share	Part blocked rate	Within-axis difference
X	4,222	38.73%	97.8%	0.3
Y	3,098	28.42%	85.5%	0.3
Z	3,580	32.84%	77.2%	23.9
Total	10,900	100.00%	86.9%	—

Distribution over the three axes. Table 6 aggregates by axis. Axis shares range from 28.4% to 38.7%, highest for X, which relates to the longer horizontal spans of the parts. There are $1{,}200 \times 6 = 7{,}200$ part-level initial mobility states, 86.9% blocked and 13.1% free; by axis, 97.8% on X, 85.5% on Y and 77.2% on Z, so interlocking is tightest along the horizontal spans. Within an axis, the blocked rates of the two opposite directions differ by less than 0.4 points for $\pm X$ and $\pm Y$, so the horizontal plane has no systematic orientation skew; Z is the only exception. This balance matters for the questions: the correct answers of the mobility and blocker types have no directional skew, models gain nothing from the marginal distribution of directions, and their accuracy can be compared directly with the random baseline.

Vertical asymmetry and its use. Z is the only exception to the symmetry above, and the asymmetry sits on the upward side: the part-level blocked rate is 65.3% for $+Z$ and 89.2% for $-Z$, while the other four directions lie between 85.4% and 98.0%. Upward blocking is concentrated on fewer parts, and most parts have no blocker along $+Z$. The skew comes from how tenon joints are built: parts are mostly inserted from above, so pulling up meets less geometric resistance than moving sideways or down. In evaluation this asymmetry is a handle rather than a defect: a model using "it can always be pulled up" as a prior will concentrate its errors on the half of instances that are blocked along $+Z$, which separates it from a model that judges from mating faces and constrained degrees of freedom; Appendix F measures exactly this prior.

K Verification Details per Question Type

Label hosts and question scope. Geometric labels have three hosts: part labels (color, mobility and blockers per direction), pair labels (contact, forced precedence) and whole-structure labels (most-connected part, initially removable set, number of valid orders). The host is not the scope of Section 3.1: the scope of a question is set by how many parts must be inspected to answer it, so "which part blocks this one along $-X$" has a part-level host and scope L2. All labels are programmatically exact; the only exception is in scoring, not labeling: full-sequence questions are scored by feasibility replay.

Thresholds and skip conditions per type. Contact is decided by a small translation of each part along each of the six directions, a boolean intersection and a threshold on the overlap volume. Contact count and the most-connected part are read from the degrees of the contact graph; the latter is skipped on ties. Color questions use a visibility guard: for every colored pixel of the colored render, hue is compared only among the parts whose silhouettes cover that pixel, so color attribution is limited to parts that actually overlap it; occluded parts get no question and their colors are not used as distractors. Without this spatial constraint, nearest-hue assignment swaps two similar palette rows wholesale and the occlusion decision fails. Directional mobility is decided by offsetting the part by a fixed fraction of its extent in that direction and comparing the total overlap with the other parts before and after; unique-free-direction questions are generated only when exactly one of the six directions is free. The blocker is the part whose overlap grows after the offset; only instances with a unique blocker are used.

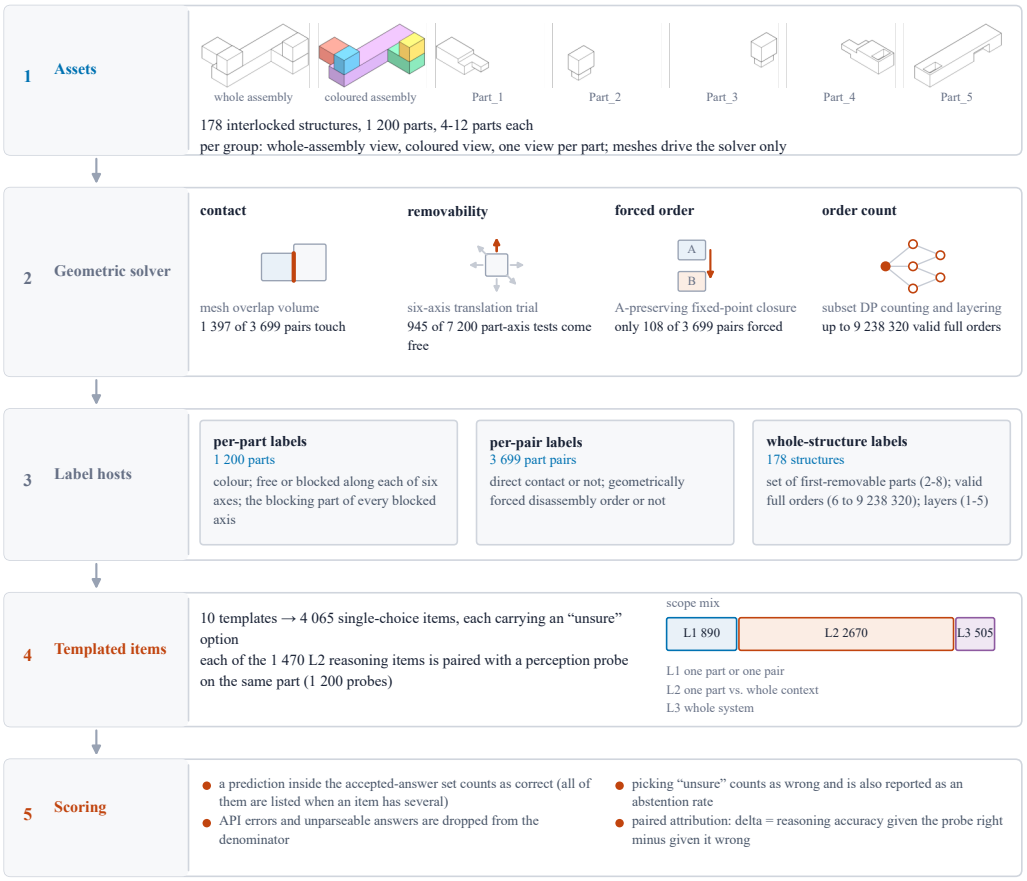

Figure 15: The five stages of question-bank construction: assets and renders; the geometric solver giving contact, directional mobility, forced precedence and order counts; labels grouped by host into part, part-pair and whole-structure labels; templated question generation with a perception probe paired to every L2 reasoning question; scoring and paired attribution. All numbers are measured on the evaluation set (178 assemblies).

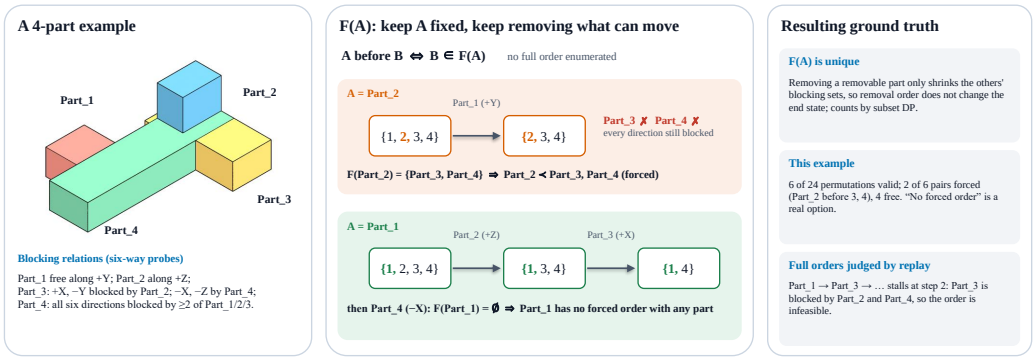

Figure 16: Enumeration-free decision of forced precedence on a 4-part structure. Left: blocking relations. Middle: keep A fixed and repeatedly remove removable parts; the remaining set F(A) is the set of parts that must come after A; Part_2 yields two forced pairs and Part_1 none. Right: the resulting partial order, order count and replay judgment of a full order.

Cross-validation of ground truth. The closure solver was compared with an implementation that enumerates all valid orders: on the 169 assemblies where enumeration stayed within its limit, forced

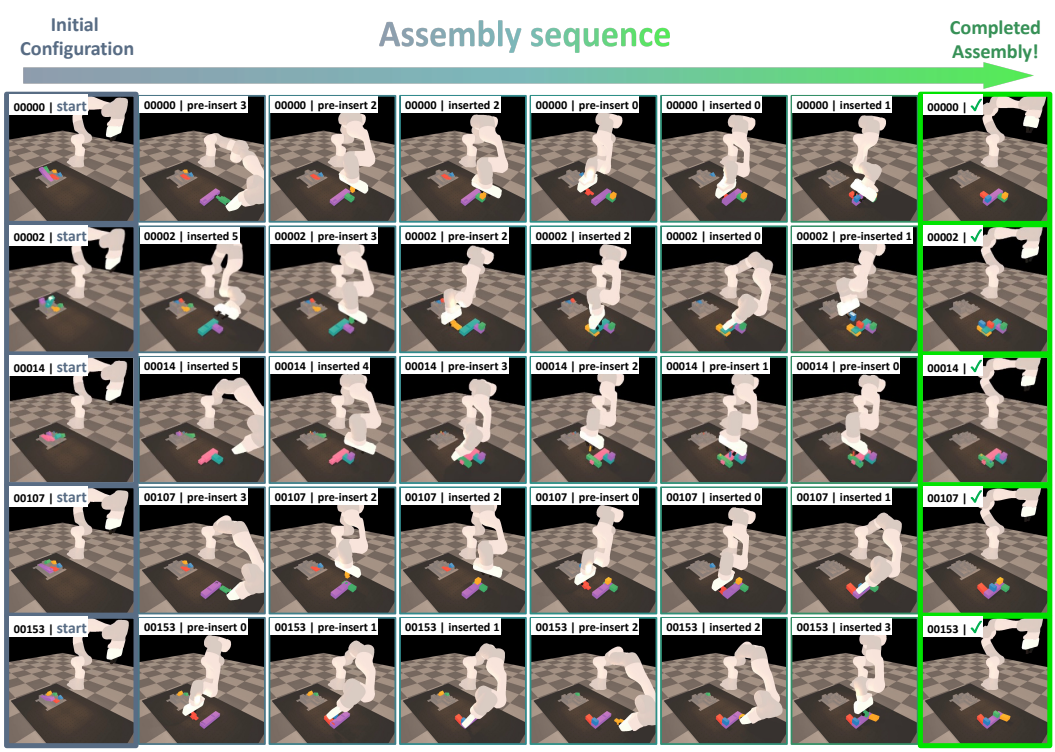

Figure 17: Reassembly of 5 structures by a robot arm in simulation, one row per structure (structure ID at the left): from left to right, the loose parts, pre-insertion and insertion states part by part, and the completed assembly. The part order comes from the solver of Section 3.3; execution translates each part along its feasible direction only.

precedence, the initially removable set and the order count agree on every assembly; the other 9 assemblies exceed the enumeration limit (the largest has 9,238,320 orders), and their exact counts are available only from the closure and the dynamic program. All decisions are purely geometric, without grasping or collision margins; as a complementary check, 5 assemblies were reassembled by a robot arm in simulation following the solver's order (Appendix L); this check covers those 5 assemblies only and is not a general claim about all labels or real-robot executability.

Full-sequence questions were also audited for text-only shortcuts: using only the text of the four candidate orders, can a heuristic beat the 25% of a four-way choice? On the 177 questions, picking the order with the most inversions hits 12.4%, the fewest inversions 23.7%, the order most different from the other three 22.6%, none above random; the label falls in the four positions 21.5%, 25.4%, 25.4%, 27.7% of the time. The only cue above random is a unique last part: in 56 questions (31.6%) exactly one order ends with a part that the other three do not, and that order is the infeasible one 46.4% of the time (14.7% over all questions); the next version will match the last part when generating valid distractor orders.

L Reassembly in Simulation

The decisions above are purely geometric and contain no grasping or collision margins. To confirm that they hold at the execution level, we reassembled 5 structures with a robot arm in simulation: a single arm grasps each part in the solver's order, moves it to a pre-insertion pose and inserts it along that part's feasible direction; all 5 were assembled to completion (Figure 17). The feasible orders given by the labels are thus not only geometrically valid but executable as action sequences; the check covers these 5 assemblies only and is not a general claim about all labels.

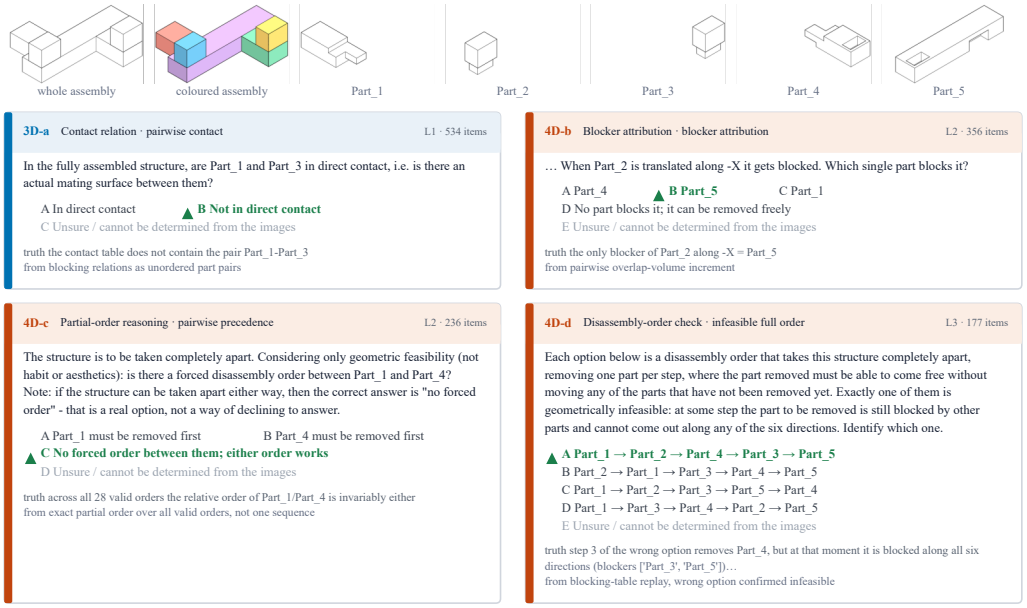

Figure 18: Four questions from one 5-part structure, with its input images above: static contact (3D-a), blocker attribution (4D-b), pairwise precedence (4D-c) and full sequence (4D-d). Question and options are the model input verbatim (fixed wording repeated across questions is elided as "…"); the green triangle marks the reference answer; the last two lines of each panel give the ground truth and how it was computed. The 4D-c prompt contains the sentence "no forced order is also a real option", which is where partial-order ground truth differs from sequence matching.

M Example Questions per Type

One question of each of the ten types of Section 3.1, all on the same structure, is shown in Figure 19; Figure 18 enlarges four of them.

N Protocol for Correlation with Embodied Downstream Tasks

If benchmark scores are to predict a model's performance in real manipulation, an external reference is needed. Our procedure: take a set of models with public scores on an embodied-manipulation benchmark, re-evaluate them on this benchmark under one protocol, compute rank correlations between the two sides, and declare the reading of each possible outcome before the analysis.

External side. We use three subsets of EmbodiedBench (Yang et al., 2025): EB-Navigation (spatial-relation judgment in long-horizon navigation), EB-Manipulation (success rate of tabletop pick-and-place) and EB-Habitat (indoor interaction). EB-Manipulation is most directly related: its failures mostly come from misjudged occlusion and contact between objects, which is the quantity our 4D dimensions isolate. Columns 2–4 of Table 7 are the public scores on these subsets, taken from the same compilation used by MMSI-Bench (Yang et al., 2026); six of the nine models are open-weight and reproducible. We use these nine rather than the 23 models evaluated here because none of the latter has public embodied scores: closed models are not integrated into frameworks like EmbodiedBench, and the evaluated versions of the open models differ from ours, so borrowing would introduce an unbounded bias. The nine models therefore need to be evaluated on this benchmark; this was not done in this version and Table 7 lists external scores only.

Benchmark side. Three quantities: overall accuracy, 4D-c (disassembly dependency) and 4D-d (full-sequence validity), the two dimensions closest to disassembly action sequences and the two lowest in Section 4.2. To control cost, evaluation uses a stratified subset of 1,001 questions rather than all 4,065: all 590 questions of 4D-c and 4D-d are kept (they are the object of analysis and not

Table 7: Models and external scores for the correlation analysis: public success rates (%) on three EmbodiedBench subsets; six of the nine models are open-weight. Correlations are to be computed over these nine points.

Model	EB-Nav	EB-Man	EB-Hab
GPT-4o	57.7	39.6	59.0
Claude-3.7-Sonnet	45.0	28.5	58.7
Claude-3.5-Sonnet	44.7	25.4	68.0
Qwen2.5-VL-72B-Instruct	40.0	16.2	37.7
InternVL2.5-78B	30.7	10.6	49.0
InternVL2.5-38B	30.3	15.8	38.3
Llama-3.2-11B-Vision-Ins	21.4	0.9	25.0
InternVL2.5-8B	21.3	7.0	11.3
Qwen2.5-VL-7B-Instruct	20.0	9.6	14.3

sampled), and 411 questions from the other dimensions are drawn proportionally within dimension $\times$ part-count bins (4–6, 7–8, 9–12 parts). Checked on the 23 evaluated models, the Spearman correlation between model rankings on the subset and on the full set is 0.994, and the largest difference in a model's accuracy between the two is 5.9 points: relative order is preserved, absolute values are not interchangeable. Dimension-level rankings are unstable at this size, so only the overall score and the 4D-c and 4D-d dimensions are used.

Readings declared in advance. Three outcomes map to three conclusions, fixed beforehand to avoid post-hoc selection. First, if 4D-c and 4D-d correlate with EB-Manipulation more strongly than the overall score does, the partial-order judgment isolated here is a component of manipulation success, and the benchmark can serve as a simulation-free proxy: it needs no arm, physics engine or episodic interaction, and one evaluation costs a batch of multiple-choice questions on static images. Second, if only the overall score correlates and the two dimensions do not, the benchmark measures the same broad ability as existing visual benchmarks and its diagnostic decomposition adds no information. Third, if neither correlates, the ability measured here is not covered by embodied-manipulation benchmarks and the two are complementary measurements. The magnitudes reported by MMSI-Bench with the same method serve as a reference: Pearson correlations of its spatial-reasoning scores with the three subsets are 0.81, 0.74, 0.73, and Spearman 0.73, 0.75, 0.62. With nine points, rank-correlation intervals are wide and conclusions can only be trends.

O Raw Rationales of a Failure Case

Accuracy and error composition do not show what the rationales themselves look like; Figure 20 gives one example of each of four failure types. This section takes one question and lists the raw rationales of four models verbatim, colored by their relation to the geometric ground truth. The question asks whether there is a forced disassembly order between parts 3 and 7 of assembly 101 (eight parts; reference answer: "no forced order", either sequence works). Ground truth: the structure has 1,112 valid disassembly orders, and both relative orders of Part_3 and Part_7 occur; Part_3 and Part_7 do not touch (Part_3 touches Part_8; Part_7 touches Part_6 and Part_8); Part_3 can be removed along $-Y$ in the initial state, and Part_7 is blocked in all six directions by Part_6 and Part_8. The two parts have no direct geometric relation; each one's mobility is decided by third parties.

Three points in Figure 21 deserve attention. First, all four models assert a forced precedence but disagree on its direction: three say Part_3 must go first, claude-opus-5 says Part_7 must. Two contradictory conclusions each come with a plausible-sounding mechanical explanation, so the coherence of a rationale is unrelated to its correctness and cannot serve as a criterion. Second, three models invent the same non-existent contact (Part_3 passing through or locking Part_7) and derive the order from it; the error begins in grounding, not reasoning—had the contact existed, "remove the locking pin before the locked part" would be correct. This also explains why strong models in Appendix G invent more than they miss: they read any two geometrically adjacent parts as interlocked by default. Third, all four models answered the paired perception probe correctly, giving the right contact count for Part_3 (the pairing of Section 4.4): they know how many parts Part_3 touches but name the wrong ones. Right count, wrong object shows that perception tested as counting does not guarantee

grounding at the relation level; the paired design separates perception from reasoning failures, but passing the perception side does not certify correct grounding.



2D-a Single-part attribute · part colour — L1 · 356 items

… Judging from the images: what colour is Part_4 in the coloured assembly image?

▲ **A emerald green** B violet C golden yellow
D coral red
E Unsure / cannot be determined from the images

truth color_map[Part_4] = emerald green
from colour table used by the renderer (part index to row)

3D-a Contact relation · pairwise contact — L1 · 534 items

In the fully assembled structure, are Part_1 and Part_3 in direct contact, i.e. is there an actual mating surface between them?

A In direct contact ▲ **B Not in direct contact**
C Unsure / cannot be determined from the images

truth the contact table does not contain the pair Part_1-Part_3
from blocking relations as unordered part pairs

3D-c Global topology / degree · contact count (paired probe) — L2 · 1 200 items

Look at the whole-assembly image. In the fully assembled structure, how many other parts are in direct contact with Part_1 (sharing an actual mating surface, not merely adjacent across a gap)?

A 0 B 4 C 2 ▲ **D 1**
E Unsure / cannot be determined from the images

truth contact count of Part_1 in the contact table = 1
from blocking relations as unordered part pairs

3D-c Global topology / degree · most-connected part — L3 · 151 items

In the final assembly, which part is in direct contact with the greatest number of other parts?

A Part_1 ▲ **B Part_5** C Part_3
D Part_4
E Unsure / cannot be determined from the images

truth contact counts {'Part_1': 1, 'Part_2': 1, 'Part_3': 1, 'Part_4': 2, 'Part_5': 3}; the unique maximum is Part_5
from degree in the contact graph

4D-a Single-part mobility · removable along a given axis — L2 · 534 items

… If Part_1 alone is translated along +Y, can it come free of the whole structure?

▲ **A It can be removed**
B It is blocked by other parts and cannot be removed
C Unsure / cannot be determined from the images

truth free directions of Part_1: ['+Y']; blockers along +Y: []
from overlap-volume solver, six-axis offsets

4D-a Single-part mobility · the one free axis — L2 · 344 items

… Among the six directions ±X / ±Y / ±Z, Part_1 can be translated out along exactly one of them. Which one?

A +X B -X ▲ **C +Y** D -Y
E +Z F -Z G None of the six directions works
H Unsure / cannot be determined from the images

truth the only free direction of Part_1 is +Y
from overlap-volume solver, six-axis offsets

4D-b Blocker attribution · blocker attribution — L2 · 356 items

… When Part_2 is translated along -X it gets blocked. Which single part blocks it?

A Part_4 ▲ **B Part_5** C Part_1
D No part blocks it; it can be removed freely
E Unsure / cannot be determined from the images

truth the only blocker of Part_2 along -X = Part_5
from pairwise overlap-volume increment

4D-c Partial-order reasoning · first-removable part — L3 · 177 items

Disassembly starts from the fully assembled structure. Without moving any other part, which of the following parts can be removed first?

A Part_5 ▲ **B Part_3** C Part_4
D Unsure / cannot be determined from the images

truth set of parts that can be removed first = ['Part_1', 'Part_2', 'Part_3']
from set of first elements over all valid orders

4D-c Partial-order reasoning · pairwise precedence — L2 · 236 items

The structure is to be taken completely apart. Considering only geometric feasibility (not habit or aesthetics): is there a forced disassembly order between Part_1 and Part_4?…

A Part_1 must be removed first B Part_4 must be removed first ▲ **C No forced order between them; either order works**
D Unsure / cannot be determined from the images

truth across all 28 valid orders the relative order of Part_1/Part_4 is invariably either
from exact partial order over all valid orders, not one sequence

4D-d Disassembly-order check · infeasible full order — L3 · 177 items

Each option below is a disassembly order that takes this structure completely apart, removing one part per step, where the part removed must be able to come free without moving any of the parts that have not been removed yet. Exactly one of them is geometrically infeasible: at some step the part to be removed is still blocked by other parts and cannot come out along any of the six directions. Identify which one.

▲ **A Part_1 → Part_2 → Part_4 → Part_3 → Part_5** B Part_2 → Part_1 → Part_3 → Part_4 → Part_5
C Part_1 → Part_2 → Part_3 → Part_5 → Part_4 D Part_1 → Part_3 → Part_4 → Part_2 → Part_5
E Unsure / cannot be determined from the images

truth step 3 of the wrong option removes Part_4, but at that moment it is blocked along all six directions (blockers ['Part_3', 'Part_5'])…
from blocking-table replay, wrong option confirmed infeasible

Figure 19: One question per type, all from the same 5-part structure. Question and options are the model input verbatim; the green triangle marks the reference answer; the last two lines of each panel give the ground truth and how it was computed. The header on the right gives the question's scope and the type's count in the whole bank (4,065 in total); blue for perception, orange for reasoning. All questions share the images at the top; each question receives only the relevant ones. 3D-b is a reserved code with no questions in this version; fixed wording repeated across questions is elided as ….

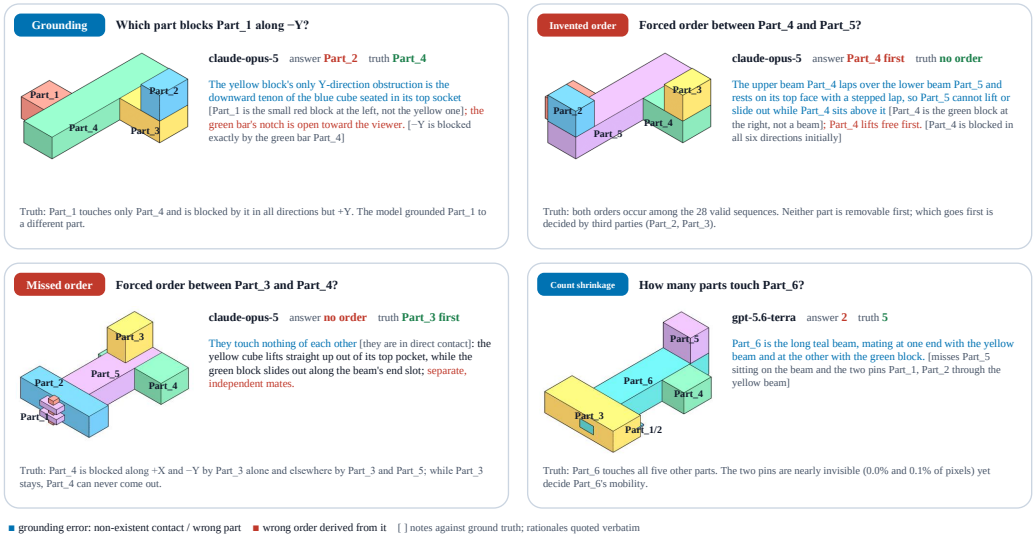

Figure 20: One example of each of four failure types, rationales quoted verbatim: a grounding error (4D-b; Part_1 in the question is grounded to a different part), an invented dependency (4D-c; a free pair declared forced), a missed dependency (4D-c; a forced precedence declared free) and contact-count shrinkage (3D-c; nearly invisible pins are not counted). Blue marks grounding errors, red the wrong order derived from them; bracketed notes compare against the ground truth.

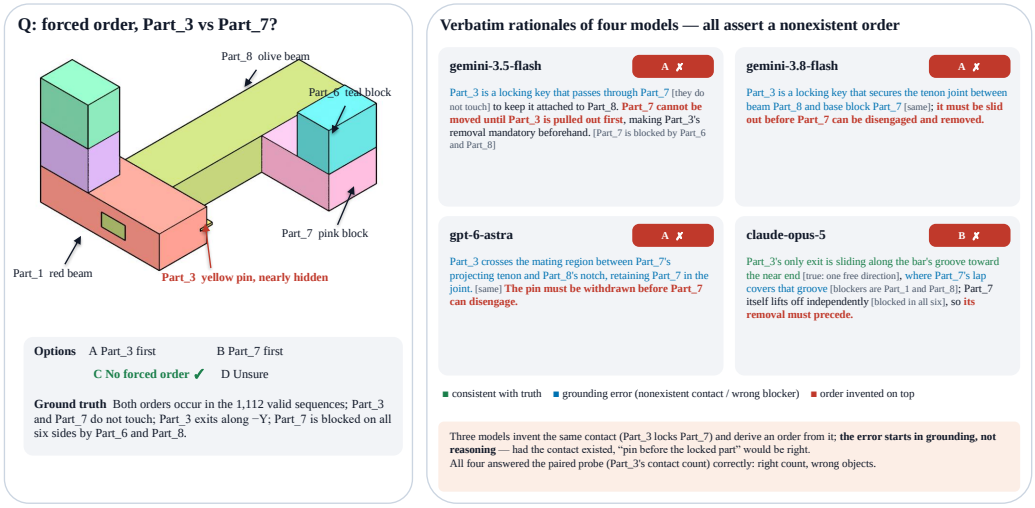

Figure 21: Raw rationales of four models on the same pairwise-precedence question, quoted verbatim. Left: the 8-part structure of the question; Part_3 is the almost hidden yellow pin; the reference answer is "no forced order". Right: green is consistent with the geometric ground truth, blue is a grounding error (a non-existent contact or a wrong blocker), red is an order invented on top of it; bracketed notes compare against the ground truth.