Spatial Amsan: A Benchmark for Perception-Grounded Spatial Reasoning and Action Evaluation in Egocentric Manipulation

Changsoo Jung[1⋆], Jack Fitzgerald[1], Ethan Seefried[1], Mariah Bradford[1], and Nathaniel Blanchard[1]

Colorado State University, Fort Collins, CO 80523, USA
Changsoo.Jung@colostate.edu

Abstract. We introduce *Spatial Amsan*, a benchmark for evaluating state-based spatial reasoning and action evaluation in egocentric manipulation videos. The term *amsan* (Korean for mental arithmetic) reflects the core challenge: constructing and updating an internal spatial state to evaluate whether each manipulation advances toward a goal. The benchmark comprises a tangram task (7 colored blocks) and a wooden puzzle task (16 blocks), totaling 282 scenarios with 1,801 evaluation turns. Models are evaluated through a multi-turn protocol that diagnostically isolates four capabilities: visual perception, temporal tracking, spatial contact reasoning, and goal-directed action evaluation. Our zero-shot evaluation of nine VLMs spanning frontier API models and open-weight local models reveals that contact reasoning is a universal bottleneck (12–16% on the wooden puzzle despite 53–96% on block identification) and that action evaluation degrades from above 79% to below 16% over 8 turns while perception remains stable. These results highlight a fundamental gap between perceiving scene changes and reasoning about their correctness over time. We release all data, annotations, and code at https://github.com/Blanchard-lab/SpatialAmsan.

Keywords: Spatial Reasoning · Vision Language Models · Egocentric Manipulation

1 Introduction

Vision-language models (VLMs) have achieved strong performance on perception and reasoning tasks [1, 16, 26], but their ability to perform structured spatial reasoning in action-oriented scenarios remains poorly understood [19, 36]. It is unclear whether current VLMs can construct and maintain an internal spatial state and use it to evaluate the correctness of sequential manipulations over time.

Most existing benchmarks evaluate spatial reasoning through single-turn relational queries such as "left of" or "above" [33]. These settings do not require models to update an internal scene representation or evaluate whether an observed

⋆ Corresponding author.

action advances toward a desired goal, providing limited insight into whether VLMs can support state-based spatial reasoning for real-world manipulation understanding.

In this work, we introduce the concept of *Spatial Amsan*. The term *amsan*, the Korean word for mental arithmetic, emphasizes internal state construction and manipulation rather than surface-level relation recognition. Spatial Amsan describes the ability to perceive a scene, abstract it into an internal spatial state, mentally modify that state, and evaluate whether a proposed action leads to success or failure under partial observability.

We propose the Spatial Amsan benchmark for perception-grounded spatial reasoning and action evaluation in egocentric video. The benchmark comprises two block manipulation tasks of increasing complexity: a tangram task with 7 colored blocks and a wooden puzzle task with 16 blocks, totaling 282 scenarios with 1,801 evaluation turns. Each task includes both successful and failed sequences, and Task 1 further introduces occlusion conditions to test spatial state maintenance under missing visual information.

Models are evaluated through a multi-turn question-answering protocol that probes four capabilities: visual perception (Q1), temporal state tracking (Q2), fine-grained spatial contact reasoning (Q3), and goal-directed action evaluation (Q4), with a conditional follow-up (Q4-2) testing reversibility reasoning. We evaluate nine VLMs spanning frontier API models (GPT-5.2 [23], Claude Sonnet 4.6 [2], Gemini 3 Flash Preview [11]) and open-weight local models (Qwen2.5-VL-7B [4], Qwen3-VL-8B [3], InternVL3-8B [39], InternVL3.5-8B [34], LLaVA-OneVision-7B [15], VideoChat-R1-7B [18]) in a zero-shot setting.

Our results reveal a clear capability gap. On Task 1, API models achieve 87–96% on Q1 but drop to 50–75% on Q3 and 30–51% on Q4. On Task 2, Q3 collapses to 12–16% for all API models. The most striking finding is the temporal degradation of Q4: all API models start above 79% at Turn 1 but drop below 16% by Turn 8 while Q1 remains stable, revealing a fundamental dissociation between perception and reasoning.

Several properties distinguish our benchmark from prior evaluations. Spatial Amsan requires *multi-turn state tracking* across sequential manipulations rather than single-turn relational queries. It evaluates *action correctness* through a structured error taxonomy rather than through perception alone. Occlusion is *systematically controlled* at varying turn positions. The evaluation is *diagnostically decomposed* into Q1–Q4, enabling precise identification of where and why models fail.

Our contributions are as follows:

- We formalize *Spatial Amsan* as a distinct capability for state-based spatial reasoning encompassing perception, temporal tracking, spatial relation understanding, and goal-directed action evaluation under partial observability.
- We introduce an egocentric benchmark with two manipulation tasks (282 scenarios, 1,801 turns) and a multi-turn protocol that diagnostically isolates each capability through dedicated questions (Q1–Q4).

- We provide a zero-shot evaluation of nine VLMs, revealing that performance on spatial contact reasoning (Q3) is a critical bottleneck, especially on the wooden puzzle task. Furthermore, action evaluation degrades sharply over turns despite stable perception.
- We release all evaluation code, annotations, and frame data to support reproducible research.

2 Related Work

2.1 Spatial Reasoning in Vision-Language Models

Spatial reasoning has been widely studied as a core capability of vision-language models. Benchmarks such as VSR [19], What'sUp [14], and SpatialBench [35] evaluate spatial understanding through pairwise relational queries (e.g. "is A left of B?") on static images. SpatialRGPT [7] extends this to region-level reasoning with flexible descriptions. SpartQA [21] and StepGame [29] probe multi-hop spatial reasoning through textual descriptions of object arrangements. While these benchmarks effectively measure relational perception, they operate on single images and do not require models to maintain or update an internal representation of the scene across time.

Video-based benchmarks extend spatial reasoning to temporal environments. EgoSchema [20] and MVBench [17] evaluate temporal understanding in egocentric and general video respectively. Video-MME [9] provides a comprehensive multi-modal evaluation across diverse video durations. However, most video-based evaluations remain descriptive: they ask models to summarize events or answer factual questions about observed actions rather than requiring mental simulation of how the scene should evolve. Our benchmark differs by requiring models to build and update a spatial state across sequential manipulation turns, using that state to evaluate action correctness rather than simply describing what happened.

2.2 Action Understanding and Feasibility Reasoning

Action-related reasoning has been explored in both video understanding and embodied AI. Video benchmarks such as Something-Something V2 [12] and EPIC-Kitchens [8] evaluate action recognition and anticipation. Assembly101 [28] introduces procedural activity understanding in assembly tasks. In embodied AI, ALFRED [30], TEACh [24], and VIMA [13] evaluate instruction-following agents in simulated environments with access to explicit state representations and structured action spaces.

Several recent works evaluate VLMs on action plausibility. These tasks typically ask models to predict likely next actions or choose between plausible action descriptions. However, they emphasize action plausibility rather than action correctness: models are rarely required to judge whether a specific observed manipulation will succeed or fail given the current scene state and a target goal.

Our benchmark fills this gap by requiring models to evaluate each manipulation against a goal configuration and classify the specific type of error when one occurs (wrong position, wrong rotation, wrong block, or block disappearance). We also include a conditional undo feasibility question (Q4-2) to further test whether models can reason about action reversibility, a capability not addressed by existing action understanding benchmarks.

2.3 Occlusion Reasoning and Object Permanence

Occlusion and object permanence have been studied in cognitive science [31, 32] and computational settings. CATER [10] evaluates compositional reasoning about object movements including containment and occlusion. IntPhys [27] and Physion [5] test intuitive physics understanding where objects may become temporarily hidden. CLEVRER [38] evaluates causal and counterfactual reasoning about collision events in synthetic scenes. In VLM evaluation, recent benchmarks include questions about occluded or partially visible objects [25], but treat occlusion primarily as a perceptual challenge where models must recover object identity or location after temporary disappearance.

Our benchmark treats occlusion differently. Rather than testing whether models can re-identify objects after occlusion, we test whether models can maintain an accurate internal spatial state when visual information is temporarily missing and continue to perform correct action evaluation afterward. In Task 1, occluded frames are systematically inserted at varying turn positions across scenarios. The model must recognize that the scene is not visible, infer that no state change has occurred, and resume spatial reasoning in subsequent turns. This goes beyond perceptual recovery to test state-based reasoning under partial observability.

2.4 Positioning of Spatial Amsan

Prior work addresses spatial relations, action plausibility, and occlusion handling as separate capabilities evaluated in isolation [7, 22]. Spatial reasoning benchmarks test static relational queries without temporal state. Action understanding benchmarks test plausibility without spatial verification against a goal. Occlusion benchmarks test perceptual recovery without downstream reasoning. Spatial Amsan unifies these aspects into a single evaluation that requires internal spatial state construction, multi-turn state maintenance, fine-grained spatial relation understanding, and goal-directed action evaluation. Table 1 summarizes the key differences. By focusing on egocentric block manipulation under controlled conditions, our benchmark isolates the spatial reasoning capabilities that current evaluations do not fully capture.

Table 1: Comparison with representative spatial reasoning and action understanding benchmarks. *Multi-turn*: requires state tracking across sequential observations. *Action eval.*: evaluates correctness of a specific action against a goal. *Contact*: tests precise geometric contact relations. *Occlusion*: systematically introduces partial observability. *Diagnostic*: separates perception, memory, spatial relations, and reasoning into distinct scored questions.

Benchmark	Multi-turn	Action eval.	Contact	Occlusion	Diagnostic
VSR [19]					
SpatialBench [35]					
MVBench [17]					
CLEVRER [38]			✓	✓	
Assembly101 [28]	✓				
ALFRED [30]	✓	✓			
VIMA [13]	✓	✓			
Spatial Amsan (Ours)	✓	✓	✓	✓	✓

3 Methodology

3.1 Benchmark Overview

The Spatial Amsan benchmark evaluates perception-grounded spatial reasoning and action evaluation in egocentric manipulation videos. The benchmark consists of controlled block manipulation tasks recorded from a first-person viewpoint. Each task defines a target structure that must be assembled through a sequence of block placements. Models must perceive the current scene state, track manipulation history, identify spatial contact relations, and judge whether each action advances toward or deviates from the goal.

The benchmark includes two tasks with distinct block sets and structural constraints. Task 1 uses seven colored tangram blocks arranged in a flat pattern. Task 2 uses sixteen wooden puzzle blocks placed into a puzzle board. Together the two tasks comprise 282 scenarios with 1,801 evaluation turns. Fig. 1 shows the acquisition environment, and Fig. 2 presents representative Tangram and Wooden Puzzle problems spanning multiple difficulty levels.

3.2 Task Design and Scenario Construction

Each problem defines a goal configuration, and for every problem we construct multiple scenarios that include both correct and incorrect execution sequences. Incorrect scenarios introduce errors such as placing a block in the wrong position, applying an incorrect rotation, or manipulating the wrong block entirely. These errors are distributed across different stages of assembly to prevent trivial detection based on timing.

Task 1 requires assembling a flat tangram pattern from seven colored blocks (Red, Yellow, Orange, Green, Skyblue, Blue, Purple). Blocks may be placed

Fig. 1: Physical setup used for data collection. Tangram blocks (Task 1) and wooden puzzle blocks (Task 2) are placed on the table inside a small workspace enclosed by gray walls. Four LED lights are attached to the ceiling wall to provide consistent illumination.

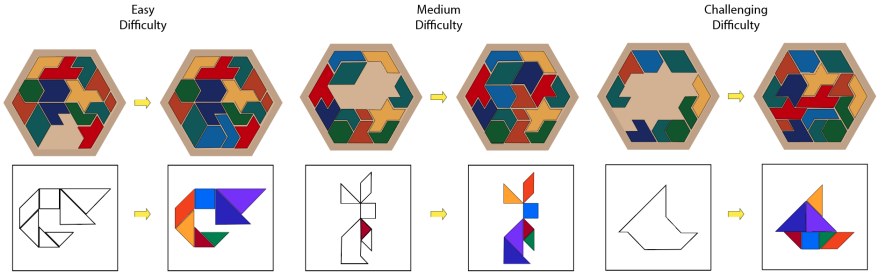

Fig. 2: Representative benchmark problems for Task 1 (Tangram) and Task 2 (Wooden Puzzle). The examples illustrate a range of structural difficulty from simpler layouts to more complex assemblies.

in any orientation including flipped configurations. The benchmark contains 15 tangram problems with 10 scenarios each, totaling 150 scenarios and 1,126 evaluation turns. Each problem is divided into visible and occluded conditions: five scenarios per problem (s01–s05) remain fully visible, and the remaining five (s06–s10) introduce a single hand-occlusion frame that temporarily obscures the view (e.g. the acting hand blocks the camera). The occlusion frame is inserted at a different turn position in each scenario to avoid positional bias. This tests whether models maintain an accurate spatial representation when visual information is briefly unavailable.

Task 2 requires placing irregularly shaped wooden blocks (B1 through B16) into a puzzle board. Both tasks are two-dimensional, but incorrect manipulations in Task 2 scenarios may involve vertical stacking as an error case. Two blocks (B3 and B4) are mirror images of each other and must be distinguished without flipping. Some blocks may already occupy their goal positions at the start of a scenario. The benchmark contains 15 puzzle problems with a variable number

of scenarios per problem, totaling 132 scenarios and 675 evaluation turns. Turns per scenario range from 2 to 10.

3.3 Annotation Protocol

Each turn's ground-truth answers for Q1–Q4 were produced by a primary annotator using a dedicated frame-by-frame annotation interface. To validate annotation reliability, we conducted a subsequent multi-annotator review with coauthors using a custom viewer that surfaces every annotation alongside the corresponding frames and flags disagreements for discussion. The review covered every annotation across both tasks and confirmed the original answers without changes, indicating that the discrete answer options effectively constrain annotator subjectivity.

3.4 Input Representation

All manipulation sequences are recorded using a stereo egocentric camera. For the experiments reported in this work, only the monocular (left) view is used. Each scenario is represented as an ordered sequence of still frames extracted at key manipulation moments. The first frame shows the goal configuration. Subsequent frames capture the state after each block placement. A separate reference image showing all labeled blocks is provided alongside each sequence. No depth maps or geometric annotations are supplied.

3.5 Evaluation Protocol

Models are evaluated through a multi-turn question-answering protocol. Turn 0 provides the task description together with the goal state image and a block reference image. Each subsequent turn presents a single manipulation frame and poses a fixed set of questions requiring JSON-formatted responses. Fig. 3 shows representative manipulation frames for Tangram and Wooden Puzzle from Turn 0 to Turn 3. In this example, an occlusion appears at Turn 2, but in the benchmark the occlusion turn varies across scenarios. The question set is identical across all turns and models. Fig. 4 shows an example of the prompt and expected response format. We describe each question and the spatial reasoning capability it targets below.

Q1 (Manipulated Block) asks which block was manipulated in the current turn. The model selects from all block labels plus "Nothing," "Not visible," and "Don't know." This question evaluates *visual perception*: the model must compare consecutive scene states and isolate the block whose position changed. Accurate block identification is a prerequisite for all downstream spatial reasoning.

Q2 (Previously Manipulated Block) asks which block was manipulated in the preceding turn. This question is introduced from Turn 2 onward and evaluates *temporal state tracking*. Since only the current frame is shown at each turn, the model must retain a working memory of prior manipulations rather than treating each frame independently.

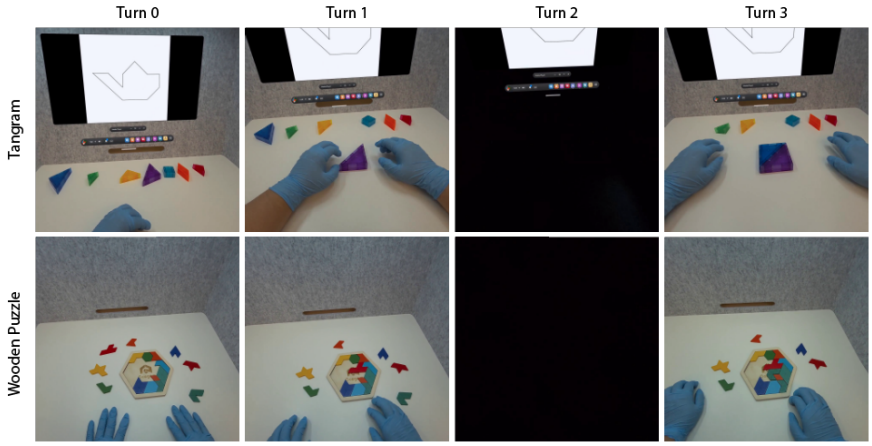

Fig. 3: Representative manipulation frames for Task 1 (Tangram) and Task 2 (Wooden Puzzle) from Turn 0 to Turn 3. Turn 0 provides the goal state image. The shown example includes an occlusion at Turn 2 for visualizations of manipulation, but the occlusion can occur at different turn positions depending on the scenario.

Q3 (Contact Blocks) asks which blocks have their side faces in contact with the manipulated block. Edge-to-edge and corner contacts do not count. In Task 2, partial contacts from overlapping blocks are included. This question evaluates *fine-grained spatial relation understanding*, requiring the model to reason about precise geometric adjacency rather than coarse proximity. Among all questions, Q3 consistently yields the lowest accuracy across models, suggesting that current VLMs struggle with exact spatial relation judgments.

Q4 (Manipulation Evaluation) asks the model to assess the correctness of the current manipulation with respect to the goal state. The ten response options span: goal reached (1), correct progress (2), correct but blocked by a past error (3), wrong position (4), wrong rotation (5), wrong block (6), block disappeared (7), not visible but reachable (8), not visible and unreachable (9), and don't know (10). A free-text reasoning field (Q4-1) accompanies every response. This question evaluates *goal-directed spatial reasoning*: the model must jointly consider the current state, the goal configuration, and the manipulation history to determine whether the action is correct or erroneous.

Q4-2 (Undo Feasibility) is a conditional follow-up triggered when Q4 indicates an error (options 4–7). The model judges whether the goal state can still be reached by undoing the current action and provides a free-text explanation (Q4-2-1). This evaluates *planning and reversibility reasoning*: whether an incorrect manipulation is recoverable or has caused irreversible damage to the assembly progress.

All prompts and question templates are fixed prior to evaluation and are fully specified in the supplementary material. No prompt tuning or task-specific adaptation is performed for any model.

Turn 0 (Setup) *[Goal Image] [Block Reference Image]*

You are presented with a sequence of egocentric-view images showing tangram manipulation. The objective is to determine whether the sequence can reach the given goal state. There are 7 tangram blocks: Red, Yellow, Orange, Green, Skyblue, Blue, Purple. [. . .] In each following turn, you will receive one image showing the current state after a manipulation.

Turn 3 *[Manipulation Frame]*

Here is Turn 3. Answer the following questions in JSON format.
Q1. Which block was manipulated in this turn? *Options:* 1=Red, . . . , 8=Nothing, 9=Not visible, 10=Don't know
Q2. Which block was the latest one manipulated before this turn? *Options:* (same)
Q3. Which block(s) are in side-face contact with the manipulated block? *Options:* (same)
Q4. Evaluation: 1. Goal reached 2. Correct 3. Correct, past error 4. Wrong position 5. Wrong rotation 6. Wrong block 7. Disappeared 8. Not visible, reachable 9. Not visible, unreachable 10. Don't know
Q4-1. Reasoning **Q4-2.** Undo feasibility (if Q4∈{4–7}) **Q4-2-1.** Reasoning

---

Expected Response

```
{"q1": 4, "q2": 6, "q3": [1, 5], "q4": [2],
"q4_1": "The Green block is correctly placed adjacent
to the goal position.", "q4_2": null, "q4_2_1": null}
```

Fig. 4: Prompt and response example for the evaluation protocol (Task 1, Turn 3). Turn 0 provides the goal image and block reference. Each subsequent turn shows one manipulation frame with questions Q1–Q4. Q2 is omitted in Turn 1. Q4-2 is conditional on error detection in Q4.

3.6 Models

We evaluate both frontier API-based models and open-weight local models in a zero-shot setting. None of the evaluated models are fine-tuned on our benchmark's data. For API models, we evaluate GPT-5.2 (OpenAI) [23], Claude Sonnet 4.6 (Anthropic) [2], and Gemini 3 Flash Preview (Google) [11] with generation parameters fixed across providers when supported by the respective APIs. For local models, we evaluate six open-weight VLMs spanning two model generations and a recent reasoning-tuned variant: Qwen2.5-VL-7B [4] and its successor Qwen3-VL-8B [3], InternVL3-8B [39] and its successor InternVL3.5-8B [34], LLaVA-OneVision-7B [15], and VideoChat-R1-7B [18], a reinforcement-fine-tuned reasoning model built on the Qwen2.5-VL backbone. All local models use bfloat16 precision with default inference settings recommended by the respective authors.

3.7 Metrics

We report per-question accuracy for Q1, Q2, Q3, Q4, and Q4-2. Overall accuracy is computed as the mean across all scored questions per turn. Q2 is excluded from Turn 1 scoring since no prior manipulation exists. Q4-2 is scored only when applicable (i.e. when the ground truth includes error categories 4–7). Results are reported per task and per model. We additionally compare performance between visible and occluded conditions in Task 1 to quantify robustness to partial observability.

4 Results

4.1 Overall Performance

Table 2 summarizes the per-question and overall accuracy for all evaluated models on both tasks. GPT-5.2 [23] achieves the highest overall accuracy on Task 1 (77.0%). Gemini 3 Flash Preview [11] reaches 74.3% and Claude Sonnet 4.6 [2] reaches 62.0%. On Task 2, Claude leads (42.8%), followed by GPT-5.2 (39.3%) and Gemini (34.7%). Among open-weight local models, InternVL3-8B [6] reaches the highest overall accuracy on both tasks (29.3% on Task 1, 17.5% on Task 2), closely followed by Qwen3-VL-8B [3] on Task 1 (28.1%). The gap between the best API model and the best local model exceeds 47% on Task 1 and 25% on Task 2, and no local model exceeds 25% on Q3 in either task.

Across all models, performance follows a consistent hierarchy among question types. Q1 (block identification) yields the highest accuracy, followed by Q2 (temporal tracking), Q4 (manipulation evaluation), and Q3 (contact reasoning). This ordering holds for both tasks and all model categories, suggesting that spatial contact reasoning is a fundamental bottleneck not resolved by scaling model size or capability.

4.2 Task 1 vs. Task 2

Task 2 is substantially harder than Task 1 for all models. GPT-5.2 drops 37.7% in overall accuracy from Task 1 to Task 2, while Gemini drops 39.6%. The largest drops occur on Q3: GPT-5.2 falls from 74.9% to 15.9% and Gemini from 69.5% to 12.2%. Two factors drive this difficulty increase. First, Task 2 uses 16 label-based blocks (B1–B16) rather than 7 color-coded blocks, making block identification harder. Second, Task 2 involves irregularly shaped blocks that produce more complex spatial contact configurations. In Task 1, 93% of turns have exactly one contact. In Task 2, turns frequently involve three or more simultaneous contacts due to irregular block shapes and tighter packing, and some error scenarios introduce overlapping placements that further increase contact complexity.

4.3 Contact Reasoning (Q3)

Q3 accuracy is the lowest among all questions for every model and task combination. To understand why, we analyze exact match versus partial overlap in Table 3. Even when exact match is low, partial overlap is significantly higher: GPT-5.2 achieves 56.9% partial overlap on Task 2 despite only 15.9% exact match. Gemini shows a similar pattern (52.2% partial vs 12.2% exact). This means models can often identify some contacting blocks but fail to enumerate the complete set.

Models also systematically over-predict "Nothing" (no contacts). In Task 2, the ground truth contains 59 "Nothing" turns, but all three API models predict "Nothing" two to three times more frequently (Claude 196, Gemini 162, GPT 158). Furthermore, models underestimate the number of contacts per turn.

Table 2: Per-question accuracy (%) on Task 1 (Tangram, 7 blocks, 2D) and Task 2 (Wooden Puzzle, 16 blocks). Q2 is excluded from Turn 1. Q4-2 is scored only on turns with error ground truth (options 4–7).

		Task 1: Tangram					
Model	Type	Q1	Q2	Q3	Q4	Q4-2	Overall
GPT-5.2 [23]	API	94.8	89.7	**74.9**	**50.5**	63.6	**77.0**
Gemini 3 Flash Preview [11]	API	**95.6**	**92.7**	69.5	42.2	68.2	74.3
Claude Sonnet 4.6 [2]	API	86.8	83.8	49.7	29.9	71.4	62.0
Qwen3-VL-8B [3]	Local	36.0	11.3	7.9	49.6	**79.2**	28.1
InternVL3.5-8B [34]	Local	23.9	9.1	24.5	6.8	56.6	17.6
VideoChat-R1-7B [18]	Local	34.1	20.4	11.8	0.7	55.7	18.0
InternVL3-8B [6]	Local	35.6	17.8	13.2	42.8	75.2	29.3
Qwen2.5-VL-7B [4]	Local	34.8	18.5	11.7	0.7	27.6	16.8
LLaVA-OV-7B [15]	Local	20.0	16.5	0.0	0.0	40.3	10.0
		Task 2: Wooden Puzzle					
Model	Type	Q1	Q2	Q3	Q4	Q4-2	Overall
Claude Sonnet 4.6 [2]	API	**65.6**	**56.3**	13.6	36.1	63.0	**42.8**
GPT-5.2 [23]	API	58.7	49.6	**15.9**	34.4	47.1	39.3
Gemini 3 Flash Preview [11]	API	52.9	45.1	12.2	28.6	56.9	34.7
Qwen3-VL-8B [3]	Local	13.5	1.1	0.1	22.1	58.2	11.1
InternVL3.5-8B [34]	Local	4.0	0.9	0.4	10.7	63.0	5.8
VideoChat-R1-7B [18]	Local	9.2	1.1	5.3	4.6	33.8	6.8
InternVL3-8B [6]	Local	6.1	7.6	1.3	**41.9**	**65.6**	17.5
LLaVA-OV-7B [15]	Local	6.1	7.9	0.0	0.0	65.6	6.9
Qwen2.5-VL-7B [4]	Local	10.7	0.7	0.1	1.9	42.9	5.0

In Task 2, 115 turns have four ground-truth contacts, but GPT-5.2 predicts four contacts only 64 times. Turns with five or six contacts are almost never predicted correctly. These patterns indicate that models default to simpler spatial configurations rather than reasoning about the full set of geometric adjacencies.

To isolate the dominant difficulty axis on Task 2, we further break down Q3 exact-match accuracy by the number of ground-truth contacts (Table 4). All three API models achieve 54–60% when a single contact is present but collapse to under 2% at two or more contacts. Local models reach at most 7.6% even at the easiest single-contact setting, and zero accuracy at higher contact counts. This pattern indicates that multi-contact geometric crowding, rather than block count alone, is the principal source of Q3 difficulty.

4.4 Temporal Degradation of Reasoning

While Q1 accuracy remains relatively stable across turn positions, Q4 accuracy degrades sharply as the number of turns increases. Table 5 shows this trend for API models on Task 1. All three models start above 79% on Q4 at Turn 1

Table 3: Q3 contact reasoning analysis for API models. Exact match requires the predicted contact set to be identical to ground truth. Partial overlap requires at least one correctly identified contact. "Nothing" counts show how often each source labels the turn as having no contacts.

Task	Model	Exact	Partial	GT Nothing	Pred Nothing
Task 1	GPT-5.2	**74.9%**	**86.1%**	327	234
Task 1	Gemini 3 Flash Preview	69.5%	82.1%	327	192
Task 1	Claude Sonnet 4.6	49.7%	76.1%	327	212
Task 2	GPT-5.2	**15.9%**	**56.9%**	59	158
Task 2	Gemini 3 Flash Preview	12.2%	52.2%	59	162
Task 2	Claude Sonnet 4.6	13.6%	37.6%	59	196

Table 4: Q3 exact-match accuracy (%) on Task 2 by number of ground-truth contacts. The "0 (Nothing)" column covers turns where the manipulated block has no contacts. API models perform well at one contact and collapse at two or more, isolating multi-contact crowding as the dominant difficulty axis.

Model	0 (Nothing)	1 contact	2 contacts	3 contacts	4+ contacts
GPT-5.2 [23]	**52.5**	**60.2**	0.0	**2.0**	0.0
Gemini 3 Flash Preview [11]	26.2	54.7	**0.8**	0.5	0.0
Claude Sonnet 4.6 [2]	39.3	55.1	0.7	1.0	0.0
Qwen3-VL-8B [3]	0.0	0.8	0.0	0.0	0.0
InternVL3.5-8B [34]	3.3	0.0	0.7	0.0	0.0
InternVL3-8B [6]	0.0	7.6	0.0	0.0	0.0
Qwen2.5-VL-7B [4]	0.0	0.8	0.0	0.0	0.0
LLaVA-OV-7B [15]	0.0	0.0	0.0	0.0	0.0

but decline severely by Turn 8: GPT-5.2 drops to 3.9%, Gemini to 15.8%, and Claude to 10.5%. In contrast, Q1 accuracy remains above 80% across all turns for all models.

This dissociation between perception and reasoning is a central finding. Models can correctly identify which block was manipulated at Turn 8 just as well as at Turn 1 but cannot determine whether the manipulation is correct. This suggests that models fail to maintain a coherent internal representation of accumulated scene state. Each turn adds to the reasoning burden, and models lack the capacity to integrate earlier observations with the current state to produce correct action evaluations.

4.5 Manipulation Evaluation Biases (Q4)

Analysis of Q4 prediction distributions reveals distinct biases across models. In Task 1, the ground truth assigns option 2 ("correct and progressing") to 535 turns and option 3 ("correct but blocked by a past error") to 325 turns. GPT-5.2 predicts option 2 for 991 turns and option 3 for zero turns, making it an

Table 5: Q1 and Q4 accuracy (%) by turn position for API models on Task 1. Q1 (perception) remains stable while Q4 (action evaluation) degrades sharply, revealing a dissociation between seeing and reasoning.

	Turn Position							
	1	2	3	4	5	6	7	8
GPT-5.2								
Q1	90.7	94.0	92.0	100.0	97.3	96.7	95.3	90.8
Q4	81.3	79.3	58.0	56.0	47.3	36.0	19.3	3.9
Gemini 3 Flash Preview								
Q1	93.3	94.7	94.7	99.3	96.7	96.0	94.0	96.1
Q4	81.3	66.0	43.3	40.7	33.3	28.0	16.0	15.8
Claude Sonnet 4.6								
Q1	89.3	85.3	81.3	92.0	89.3	84.0	89.3	80.3
Q4	79.3	35.3	28.7	22.7	24.0	20.7	8.7	10.5

overly optimistic evaluator. Claude shows the opposite bias, predicting "wrong position" 334 times and "wrong rotation" 318 times against ground-truth counts of 43 each. Gemini falls between the two extremes: it over-predicts option 2 (689) and "wrong position" (283) but also produces some option 3 predictions (22), making it less extreme than either GPT or Claude.

No model detects option 3 ("correct but blocked by a past error") reliably. This category appears in approximately 29% of Task 1 turns and 17% of Task 2 turns. GPT-5.2 never predicts it across either task, and Gemini predicts it in fewer than 2% of cases. Recognizing this category requires the model to simultaneously judge that the current action is locally correct while recalling that a previous action was incorrect and has made the goal unreachable. This combination of temporal memory and counterfactual reasoning appears to be beyond current model capabilities.

4.6 Visible vs. Occluded Conditions

Table 6 compares accuracy under visible and occluded conditions for Task 1. The largest impact of occlusion is on Q2 (temporal tracking), where GPT-5.2 drops 10.0%, Claude drops 7.8%, and Gemini drops 5.3%. Q1, Q3, and Q4 are less affected, suggesting that models can handle brief visual interruptions for perception-level tasks but struggle to maintain temporal continuity when a frame provides no information.

5 Discussion

The most consistent finding is that Q3 (contact reasoning) yields the lowest accuracy for every model and task. All three API models drop from 87–96% on

Table 6: Visible vs. occluded accuracy (%) for Task 1 API models. Occlusion primarily impacts Q2 (temporal tracking), with smaller effects on other questions.

Model	Condition	Q1	Q2	Q3	Q4
GPT-5.2	Visible	95.6	95.1	76.0	49.8
GPT-5.2	Occluded	94.2	85.1	73.8	51.2
Gemini 3 Flash Preview	Visible	95.8	95.5	68.4	41.3
Gemini 3 Flash Preview	Occluded	95.3	90.2	70.5	43.0
Claude Sonnet 4.6	Visible	89.5	88.0	48.5	28.1
Claude Sonnet 4.6	Occluded	84.3	80.2	50.8	31.5

Q1 in Task 1 to 12–16% on Q3 in Task 2. Answering Q3 requires determining whether two blocks share a side-face boundary, a geometric judgment that depends on precise boundary localization rather than coarse object detection. Models default to simpler configurations: they over-predict "Nothing" and underestimate the number of contacts when multiple exist. The gap between partial overlap and exact match (e.g. 56.9% vs 15.9% for GPT-5.2 and 52.2% vs 12.2% for Gemini on Task 2) confirms that models possess a coarse notion of proximity but lack the precision for exact geometric adjacency.

A striking dissociation appears between perception and reasoning across turn positions. Q1 remains stable at 80–100% for all three API models throughout an 8-turn sequence while Q4 drops from above 79% to below 16%. The models can still correctly identify which block was manipulated at Turn 8 but can no longer determine whether the manipulation is correct. This indicates that models process each turn largely independently rather than maintaining a coherent internal state, and the sharp Q4 decline reveals a lack of effective mechanisms for long-horizon state integration.

Q4 option 3 ("correct but blocked by a past error") exposes a particularly revealing blindspot. GPT-5.2 predicts it zero times across both tasks despite it appearing in 29% of Task 1 turns. Gemini predicts it in fewer than 2% of cases. This category requires a compound judgment: recognizing the current action as locally correct while recalling that a past mistake has already made the goal unreachable. Models can evaluate manipulations in isolation but cannot integrate historical error information into their assessments. The three API models also exhibit distinct Q4 biases: GPT over-predicts "correct" (991 vs 535 in Task 1 ground truth), Claude over-predicts errors such as "wrong position" (334 vs 43), and Gemini falls between the two extremes with moderate over-prediction on both sides.

The failure modes persist across model generations and reasoning-tuned variants. Among the six local models, InternVL3-8B and Qwen3-VL-8B reach the highest overall accuracy on Task 1 (29.3% and 28.1%). Both still trail GPT-5.2 (77.0%) by more than 47 points. The newer InternVL3.5-8B (17.6% on Task 1, 5.8% on Task 2) does not surpass InternVL3-8B. Qwen3-VL-8B improves on Qwen2.5-VL-7B (16.8%) by 11 points but remains far from API-level perfor-

mance. VideoChat-R1-7B, a reinforcement-fine-tuned reasoning model built on Qwen2.5-VL, reaches 18.0% on Task 1 and 6.8% on Task 2 (Table 2). Its Task 1 Q4 remains below 1%, tied with Qwen2.5-VL-7B and LLaVA-OV-7B as the most severe failures. Its Task 2 Q3 (5.3%) is the highest among local models but still falls to zero at two or more contacts. Q4 option 3 is also unreachable for the newer local models, with both Qwen3-VL-8B and InternVL3.5-8B predicting it zero times across either task.

5.1 Limitations

We evaluate the models only on monocular input. Evaluation with stereo views may improve depth estimation and spatial reasoning. Our benchmark covers two block manipulation tasks with distinct complexity levels but does not include deformable objects or tool use. We position deformable manipulation, tool use, and 3D assembly as future work. The controlled rigid-block scope is an intentional design choice that isolates state-based spatial reasoning from confounds such as physics dynamics, body occlusion, and manipulation skill execution. All evaluations use a zero-shot setting with fixed prompts, and chain-of-thought prompting or few-shot examples could improve individual model results. Ground-truth annotations were produced by a primary annotator and subsequently validated through a multi-annotator review with co-authors using a custom viewer that flags disagreements for discussion (Sec. 3). The discrete answer options constrain subjectivity for Q1, Q2, and Q4, and the review pass confirmed the primary annotations without changes. We also attempted the spatial supersensing video-VLM Cambrian-S-7B [37], but its video-frame-centric multi-image preprocessing is incompatible with our three-context-image protocol per turn.

6 Conclusion

We introduced Spatial Amsan, a benchmark for evaluating perception-grounded spatial reasoning in egocentric manipulation videos. The benchmark comprises two tasks (282 scenarios, 1,801 turns) with a diagnostic protocol that separates perception (Q1), temporal tracking (Q2), spatial contact reasoning (Q3), and goal-directed action evaluation (Q4). Our evaluation of nine VLMs (three API and six local) reveals that spatial contact reasoning is the universal bottleneck (all API models score 12–16% on Q3 in the wooden puzzle), action evaluation degrades sharply over turns while perception remains stable (Q4 drops from above 79% to below 16% across 8 turns for all API models), and no model can detect when correct actions are undermined by past errors. These findings demonstrate that current VLMs lack the state-based spatial reasoning required for manipulation understanding, motivating architectures that explicitly maintain spatial state across sequential observations. We release all data, annotations, and evaluation code to support future research.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Anthropic: System card: Claude opus 4 & claude sonnet 4. Tech. rep., Anthropic (2025), `https://www.anthropic.com/claude-4-system-card`, accessed: 2026-03-05
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Bear, D., Wang, E., Mrowca, D., Binder, F., Tung, H.Y., Pramod, R.T., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., Li, F.F., Kanwisher, N., Tenenbaum, J.B., Yamins, D.L.K., Fan, J.E.: Physion: Evaluating physical prediction from vision in humans and machines. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2021)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: SpatialRGPT: Grounded spatial reasoning in vision-language models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37 (2024)
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The EPIC-KITCHENS dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 753–771 (2018)
9. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24108–24118 (2025)
10. Girdhar, R., Ramanan, D.: CATER: A diagnostic dataset for compositional actions and TEmporal reasoning. In: International Conference on Learning Representations (ICLR) (2020)
11. Google: Gemini 3 flash: Our fastest, most efficient thinking model. Google Blog (2025), `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`, accessed: 2026-03-05
12. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5843–5851 (2017)
13. Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., Fan, L.: VIMA: Robot manipulation with multimodal prompts. In: Proceedings of the International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 14975–15022 (2023)

14. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9161–9175 (2023)
15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
16. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
17. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: MVBench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22195–22206 (2024)
18. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
19. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
20. Mangalam, K., Akshulakov, R., Malik, J.: EgoSchema: A diagnostic benchmark for very long-form video language understanding. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 46212–46244 (2023)
21. Mirzaee, R., Rajaby Faghihi, H., Ning, Q., Kordjamshidi, P.: SPARTQA: A textual question answering benchmark for spatial reasoning. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). pp. 4582–4598 (2021)
22. Mittal, H., Agarwal, N., Lo, S.Y., Lee, K.: Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18580–18590 (2024)
23. OpenAI: Gpt-5 system card. Tech. rep., OpenAI (2025), `https://arxiv.org/abs/2601.03267`, accessed: 2026-03-05
24. Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., Hakkani-Tur, D.: TEACh: Task-driven embodied agents that chat. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2017–2025 (2022)
25. Pothiraj, A., Stengel-Eskin, E., Cho, J., Bansal, M.: Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8001–8010 (2025)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Riochet, R., Castro, M.Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., Dupoux, E.: IntPhys 2019: A benchmark for visual intuitive physics understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5016–5025 (2022)
28. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural

activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21096–21106 (2022)
29. Shi, Z., Zhang, Q., Lipani, A.: StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11321–11329 (2022)
30. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10737–10746 (2020)
31. Spelke, E.S.: Principles of object perception. Cognitive Science **14**(1), 29–56 (1990)
32. Tang, H., Kreiman, G.: Recognition of occluded objects. In: Computational and cognitive neuroscience of vision, pp. 41–58. Springer (2016)
33. Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., Li, S., Joshi, N.: Is a picture worth a thousand words? delving into spatial reasoning for vision language models. Advances in Neural Information Processing Systems **37**, 75392–75421 (2024)
34. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
35. Xu, P., Wang, S., Zhu, Y., Li, J., Qi, G., Zhang, Y.: SpatialBench: Benchmarking multimodal large language models for spatial cognition. arXiv preprint arXiv:2511.21471 (2025)
36. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
37. Yang, S., Yang, J., Huang, P., Brown, E., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
38. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: Collision events for video representation and reasoning. In: International Conference on Learning Representations (ICLR) (2020)
39. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)