REMAP: EVALUATING GEOMETRIC DUAL REPRESENTATIONS IN MULTI-VIEW SPATIAL REASONING

Selina Cheng[1], **Anne Wu**[2], **Eunice Yiu**[3], **Yoav Artzi**[2]
[1]Hunter College and Macaulay Honors College, City University of New York
[2]Department of Computer Science, Cornell University
[3]Department of Psychology, University of California, Berkeley

ABSTRACT

Building coherent world models requires agents to align local perceptual experience with global, abstract representations of space. This paper introduces REMAP, a controlled benchmark for evaluating multiview spatial reasoning in vision-language models (VLMs). Motivated by developmental research showing that humans flexibly align egocentric and allocentric representations across changes in viewpoint and orientation, the task requires agents to identify a target location by aligning an allocentric map with multiple egocentric observations. Critically, this setup tests cross-view geometric correspondence rather than view-specific visual matching. REMAP instantiates synthetic triangle environments with systematically varied angle configurations, enabling fine-grained analysis of sensitivity to different geometric relations and representations. Evaluating 17 VLMs alongside human performance, we find that leading models outperform random baselines but remain substantially below average human accuracy. Beyond this performance gap, models exhibit systematic, representation-specific failures: they show persistent weaknesses on side-based representations even when geometric cues are highly distinctive. These findings reveal a substantial gap between human and model spatial reasoning, suggesting that current VLMs lack the cross-view geometric abstractions and struggle to robustly integrate partial observations needed for coherent world model construction.

1 INTRODUCTION

"Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her." —Lewis Carroll, *Alice's Adventures in Wonderland*

Humans maintain a coherent sense of space even as viewpoint, scale, and orientation change. Like Alice falling through the rabbit hole, we track spatial relations across continuous transformation, constantly updating where we are *relative* to the environment. This abstract understanding of space goes beyond recognizing objects or scenes from familiar views. This ability supports robust navigation in unseen environments and enables efficient, geometry-aware reasoning across changes in viewpoint and representation.

Decades of developmental research show that spatial reasoning relies on abstract geometric representations rather than visual recognition alone (Landau et al., 1981; Izard & Spelke, 2009). These geometric representations support success in symbolic spatial tasks, but are not initially unified (DeLoache, 1991; 2000). Children flexibly recruit different geometric relations depending on task demands, with little evidence of fully integrated Euclidean reasoning early in development (Dillon et al., 2013; Dillon & Spelke, 2018). By the preschool years, however, symbolic artifacts such as maps scaffold alignment between these representations: map-based tasks reveal systematic patterns of success and difficulty that reflect which geometric relations are informative, providing a principled and well-characterized benchmark for testing abstract spatial representations. Such tasks therefore offer a natural starting point for evaluating whether VLMs can align allocentric spatial descriptions with egocentric visual views in a geometry-sensitive manner.

These findings raise a fundamental question for artificial intelligence. VLMs have been tested on increasingly rich spatial reasoning benchmarks, spanning visual simulations, perspective reasoning, and large-scale spatial cognition (Chen et al., 2024; Li et al., 2026; Ramakrishnan et al., 2024). However, it remains unclear whether success on these tasks reflects abstract, geometry-based spatial representations or reliance on view-specific visual cues. Map-to-view alignment tasks directly probe this distinction by requiring symbolic alignment between allocentric and egocentric representations. In this work, we introduce a controlled spatial reasoning task, grounded in human developmental paradigms, to test this capability in modern VLMs, and whether their limitations resemble signatures of human symbolic spatial reasoning.

2 Related Work

2.1 Cognitive Foundations of Spatial Reasoning

Research in cognitive science characterizes human spatial reasoning as grounded in early-emerging geometric representations that support navigation, object perception, and symbolic reasoning. Infants, young children, and non-human animals reliably use distance and directional relations of extended surfaces to reorient in space, often independently of landmarks or object identity (Landau et al., 1981; Lee et al., 2012). In contrast, visual form analysis relies on angle, relative length, and shape relations that are invariant to scale and orientation (Izard & Spelke, 2009).

Crucially, these geometric systems are initially dissociated. A substantial body of work shows that children recruit different geometric relations depending on task demands, rather than reasoning within a single unified Euclidean framework. For example, Dillon et al. demonstrate that when interpreting the same spatial symbol (e.g., a map), children selectively rely on distance or angle information depending on which relations are relevant for successful performance (Dillon et al., 2013). Despite identical stimuli and instructions, children's errors reveal systematic sensitivity to specific geometric dimensions, with little evidence of integrated spatial representations early in development (Dillon & Spelke, 2018). Related work on symbolic development further shows that successful spatial reasoning requires treating symbols (e.g., maps or scale models) as representations of larger environments, a coordination that is cognitively demanding even when the depicted objects are perceptually familiar (DeLoache, 1991; 2000).

Together, these findings suggest that spatial reasoning depends on the ability to flexibly align abstract geometric representations across perspectives, rather than on perceptual matching alone. This view provides a cognitive foundation for studying how human or artificial agents relate symbolic spatial descriptions to egocentric visual experience.

2.2 Evaluation of multimodal spatial reasoning

Recent VLMs show improving performance on spatial benchmarks, but it is still debated whether this reflects genuine geometric abstraction or shortcutting.

One line of work attempts to provide VLMs with stronger spatial capabilities via architectural or training interventions, e.g., by incorporating explicit spatial representations or supervision targeted at geometry and perspective (Chen et al., 2024). A complementary line evaluates whether spatial cognition emerges in frontier multimodal models without task-specific adaptation, using suites of tasks that span perspective taking, navigation-style reasoning, and large-scale simulated environments (Ramakrishnan et al., 2024; Li et al., 2026). These evaluations have been useful for measuring broad competence, but many settings inevitably mix geometry with semantic recognition (e.g., object identity, textures, language priors) and environment-specific cues, making it difficult to attribute success or failure to the ability to draw on and reason over specific geometric relations.

A growing number of recent benchmarks focus instead on mental model construction from limited views: given a small number of egocentric observations, models must infer a coherent spatial configuration that supports downstream queries (e.g., localization, relative position, or layout inference) (Yin et al., 2025; Zhang et al., 2026). In parallel, representation-level analyses probe whether VLMs contain disentangled spatial signals and how such signals vary across layers, training regimes, and visual backbones (Yang et al., 2025). Together, these efforts motivate evaluations that separate (i)

view-specific matching from (ii) cross-view geometric correspondence, and that diagnose which geometric cues a model actually uses during inference.

Our benchmark complements prior work by isolating this distinction in a controlled, psychology-inspired map-to-view alignment task. By pairing an allocentric symbolic map with multiple egocentric views and systematically manipulating both triangle angle configuration and representational format (FULL/CORNERS/SIDES), REMAP targets a core requirement of world modeling: aligning information across coordinate frames under partial viewpoint change, while being sensitive to the underlying geometry. This design enables fine-grained error analyses (e.g., distance-/angle-driven confusions and failures to integrate edge information) that are hard to localize in richer, semantically grounded environments. It directly connects modern VLM evaluation to well-characterized signatures from human spatial reasoning paradigms (Dillon et al., 2013; Dillon & Spelke, 2018).

3 REMAP: THE REPRESENTATIONAL MAPPING BENCHMARK

To isolate geometric intuition from the confounds of high-level visual recognition (e.g., rich semantic landmarks such as furniture, textures, or color cues), we introduce REMAP, a minimal, synthetic benchmark grounded in classic developmental paradigms of map-based geometric reasoning (Dillon et al., 2013; Dillon & Spelke, 2018). REMAP evaluates cross-view geometric correspondence: an agent must align local egocentric observations with a global allocentric map under changes in viewpoint, rather than solving the task by view-specific visual matching (Figure 1).

Task. Figure 1d provides an example. Each problem instance provides an allocentric overhead map of a triangular enclosure with a target location marked by a yellow circle, together with three egocentric first-person views captured from inside the same enclosure. In the egocentric views, visible candidate locations are annotated with integer labels. Within a question, the same integer refers to the same physical location across all three views, and exactly one integer corresponds to the target location marked on the map. The task is to predict the correct integer given the map and the three views, and we evaluate performance using accuracy. Numeric labels vary across questions while preserving within-question consistency. We additionally construct questions so that the target location in the map is visible in at least one of the three egocentric views, ensuring that each instance is answerable from the provided evidence.

Controlled factors. We systematically manipulate four aspects of the benchmark.

- First, we vary **triangle type** using four angle configurations: 50-60-70, 40-60-80, 35-60-85, and 30-60-90 (Figure 1b). These configurations produce different levels of geometric discriminability: triangles with more similar internal angles require finer-grained alignment between egocentric and allocentric geometry.
- Second, we vary the **representational format** of the enclosure (Figure 1c): FULL encloses the full perimeter, CORNERS places walls at each angle, and SIDES places wall segments along the middle of each edge.
- Third, we vary the **target type** over six canonical target locations (Figure 1a) defined relative to the triangle's angles: locations at each of the smallest, medium, and largest angles, and locations opposite each of those angles at the midpoint of each side length.
- Fourth, we vary **viewpoint**: egocentric images are rendered from a camera that rotates in 30° yaw increments, and each question uses a triplet of three consecutive yaw views, which may be displayed in a permuted order.

Rendering and dataset instantiation. Both the allocentric maps and egocentric views are rendered in Blender (Blender Foundation, 2025), with maps obtained via top-down rendering, and egocentric views captured from within the triangular enclosure. Table 1 summarizes the statistics of the resulting benchmark. Across triangle types, we render 18 map images per configuration (3 representational formats $\times$ 6 target types) and 108 egocentric images per configuration (3 formats $\times$ 12 viewpoint triplets $\times$ 3 views). Questions are formed by pairing maps with viewpoint triplets, and retaining only those pairs for which the target appears in at least one of the three egocentric views, yielding a total of 519 questions in REMAP.

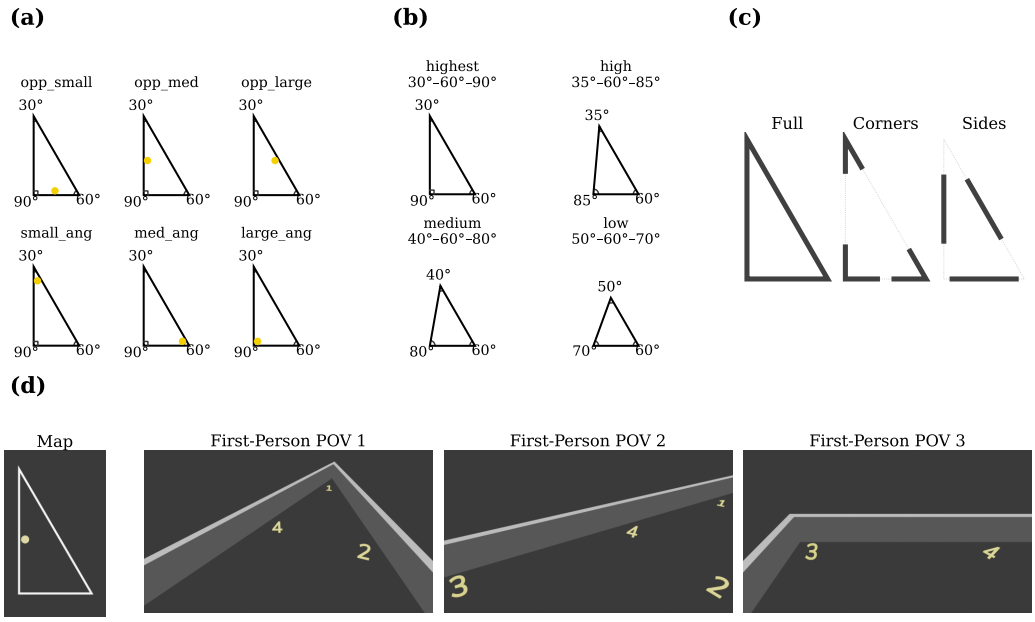

Figure 1: Overview of the design of REMAP. (a) Six target locations adapted from classic developmental studies of geometric reasoning (Dillon et al., 2013; Dillon & Spelke, 2018) shown with the 30-60-90 triangle. (b) Four triangle angle configurations. (c) Three wall configurations shown for the 30-60-90 triangle: FULL encloses the full perimeter, CORNERS places walls of equal lengths only at the angles, SIDES places wall segments along the middle of each edge. Dashed lines indicate the full triangle outline for reference. (d) An example of a task. Participants view an overhead map (left) showing a target location (yellow disk) and three consecutive first-person (egocentric) views (right) taken from inside the enclosure. The task is to identify the target's position on the map given the numbered landmarks visible in the egocentric views. In this example using a full 30-60-90 triangle, the model or participant is asked to locate the number corresponding with the yellow disk (opposite the 60° angle). In this example, the correct answer is "4".

Split	Angles (°)	Map Images	POV Images	Qs	Q/setup
Low	(50,60,70)	18	108	123	41
Medium	(40,60,80)	18	108	129	43
High	(35,60,85)	18	108	135	45
Highest	(30,60,90)	18	108	132	44
Total	–	72	432	519	43.25

Table 1: Statistics of the REMAP benchmark. For each of the four triangle angle configurations, we report the corresponding angles, the counts of generated overhead map and first-person view images, the count of valid questions pairing one map with three first-person views, and the mean number of questions per representational format ('Q/setup'). Valid questions are those in which the target location marked on the map is visible in at least one of the three first-person views.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation Metrics. For the spatial reasoning task, we report accuracy for both the model and humans. We also report reaction times for human performance.

Models. To cover a variety of VLMs, we consider several families of state-of-the-art models: (1) for closed-source models, we use GPT 5.2 (OpenAI, 2025), Gemini 2.5 Flash, Gemini 3 Flash and

Gemini 3 Pro (Google DeepMind, 2025), Gemini Robotics ER-1.5 (Google DeepMind Robotics, 2025) and Claude Sonnet 4.5 (Anthropic, 2025); (2) for open-source models, we experimented with multiple sizes of Qwen3-VL (Team, 2025) and InternVL (Wang et al., 2025). For the closed-source models, we used their default setup. Gemini 3 uses dynamic reasoning by default. For GPT 5.2, we test both the default configuration and a reasoning effort set to medium. We use a temperature of 2 for Gemini and 1 for Claude. We set the maximum output length to 65,536 tokens for all closed-source models, except for Claude where the maximum is 64,000. For Qwen3-VL and InternVL-3.5, we set it to 32,768 tokens. We conduct experiments across 5 random seeds.

Human evaluation. To assess human performance on the REMAP benchmark, we recruited 41 participants between the ages of 18 and 32. The questions were randomized and distributed such that each instance in the benchmark was completed by 1-2 human participants.

4.2 Main results

Table 2: Model performance across all four angle configurations and three triangle representations. For each angle configuration, we report average accuracy as mean$_{\pm\text{std}}$ over 5 runs (seeds). The right block reports the accuracy per representational format (Full/Corners/Sides) averaged over the four angle configurations. In the left block, **green** and light green highlight the best and second-best model per angle configuration (column). In the right block, **blue** and light blue highlight the best and second-best triangle representation per model, while pink marks the worst. Table 5 in the Appendix features the full results across all models.

	Avg (A) by angle configuration (%)					Avg by rep.		
Model	50-60-70	40-60-80	35-60-85	30-60-90	A	F	C	S
Random	29.3	27.9	26.7	27.3	27.8	27.8		
Human (average)	74.0	74.2	83.0	75.6	76.7	70.6	83.3	75.4
Closed-Source Models								
Gemini 3 Pro	48.3$_{\pm 2.2}$	**54.3$_{\pm 3.3}$**	**54.8$_{\pm 1.7}$**	**55.8$_{\pm 1.1}$**	**53.3**	56.2	55.3	48.3
Gemini 3 Flash	**53.2$_{\pm 0.9}$**	53.3$_{\pm 1.7}$	50.8$_{\pm 1.1}$	51.1$_{\pm 3.2}$	52.1	49.9	58.1	48.3
Gemini Robotics ER-1.5	44.7$_{\pm 2.0}$	43.6$_{\pm 2.4}$	41.2$_{\pm 3.3}$	40.8$_{\pm 1.6}$	42.6	41.7	52.0	34.0
GPT 5.2 (Medium R.)†	46.7$_{\pm 3.2}$	49.0$_{\pm 3.1}$	50.4$_{\pm 1.9}$	52.1$_{\pm 2.3}$	49.5	48.7	57.4	42.5
GPT 5.2	43.3$_{\pm 2.6}$	43.4$_{\pm 3.2}$	47.0$_{\pm 2.6}$	46.1$_{\pm 2.2}$	44.9	45.5	50.7	38.5
Claude Sonnet 4.5	42.4$_{\pm 1.6}$	38.8$_{\pm 1.6}$	39.0$_{\pm 1.0}$	35.6$_{\pm 0.9}$	38.9	39.3	36.1	41.5
Open-Source Models								
Qwen3-VL-2B	32.7$_{\pm 3.7}$	25.7$_{\pm 4.6}$	30.2$_{\pm 2.8}$	28.5$_{\pm 4.2}$	29.3	30.8	26.7	30.4
Qwen3-VL-4B	35.0$_{\pm 1.5}$	31.9$_{\pm 3.5}$	33.0$_{\pm 2.8}$	38.0$_{\pm 1.8}$	34.5	42.1	32.8	28.5
Qwen3-VL-32B	40.7$_{\pm 4.5}$	38.4$_{\pm 1.4}$	38.4$_{\pm 1.0}$	37.6$_{\pm 2.0}$	38.8	42.0	41.4	32.9
InternVL-3.5-2B	30.6$_{\pm 1.6}$	30.2$_{\pm 3.1}$	31.0$_{\pm 3.0}$	33.5$_{\pm 1.5}$	31.3	33.5	29.1	31.3
InternVL-3.5-14B	31.5$_{\pm 0.9}$	34.4$_{\pm 2.1}$	36.3$_{\pm 1.4}$	35.2$_{\pm 1.6}$	34.4	34.2	30.6	38.2
InternVL-3.5-38B	34.1$_{\pm 1.4}$	34.7$_{\pm 1.0}$	31.3$_{\pm 2.4}$	31.7$_{\pm 2.4}$	33.0	37.7	30.8	30.3

†GPT 5.2 with medium reasoning effort.

1. All evaluated VLMs substantially underperform humans. Table 2 reports performance across the four triangle angle configurations. Averaged over configurations and representations in the REMAP benchmark, human accuracy reached 76.7% on average, compared to 53.3% for the strongest model (Gemini 3 Pro) and 27.8% for the random baseline. The best-performing closed-source models (Gemini 3 Pro and Gemini 3 Flash) achieved 53.3% and 52.1% respectively, closely followed by GPT 5.2 with medium reasoning effort, with 49.5%. GPT 5.2, Gemini Robotics ER-1.5, and Claude Sonnet 4.5 clustered between 39% and 45%. Among open-source models, performance remained near chance for smaller variants, with Qwen3-VL-32B (38.8%) approaching but not exceeding the weakest closed-source model. Despite non-trivial above-chance performance, a

Category	Mean ± Std. Error
FULL	17.94 ± 1.49
CORNERS	19.47 ± 1.47
SIDES	22.22 ± 1.48

Table 3: Summary statistics of human reaction time by representation format (mean ± standard error).

substantial gap remains between VLMs and human accuracy, indicating that the task is tractable for humans yet not solved by current frontier models.

2. Representation format affects humans and VLMs differently. Averaged across triangle types, human accuracy varied systematically by representation format. Accuracy was highest in CORNERS (83.3%), followed by SIDES (75.4%) and FULL (70.6%). A mixed-effects logistic regression confirmed that CORNERS significantly outperformed FULL (p=.003). In contrast, frontier VLMs displayed a qualitatively different profile. For nearly all closed-source models except Claude, SIDES was the lowest-performing representation even when angles were highly separable. Claude showed an inverted pattern, performing best in SIDES and worst in CORNERS. Open-source models exhibited weaker and less consistent separation across representation formats. Whereas humans show a clear advantage for corner-based representations, VLMs do not exhibit a stable or human-aligned representation preference.

3. Triangle type modulates human but not model representation sensitivity. Human representation sensitivity depended on how discriminable the internal angles in the triangle were. In the least separable triangle geometry (50-60-70), overall accuracy was lower: FULL yielded the highest accuracy (78.1%), with CORNERS (75.0%) and SIDES (68.4%) trailing. In more separable triangle geometries (40-60-80/35-60-85/30-60-90), humans achieved consistently high accuracy in CORNERS (83%–88%), typically outperforming both FULL and SIDES. Most VLMs did not show this adaptive pattern. Their weaker performance in SIDES persisted even in highly separable geometries. For example, Gemini 3 Pro scored 60.5%/59.6% in CORNERS versus 48.6%/49.8% in SIDES for 30-60-90/35-60-85 respectively (Table 5 in the Appendix). This suggests that models are limited in leveraging side-based geometric constraints rather than merely sensitivity to angle discriminability.

4. Corner representations facilitate more efficient human reasoning. Mean human response times ranged from 17-23 s across representations. In Table 3, results are computed after removing ±2 SD on both overall accuracy and mean reaction time. SIDES representations yielded the slowest responses (22.22 s), while CORNERS achieved higher accuracy without significant increase in response time compared to FULL. Mixed-effects regression confirmed that SIDES was 24.5% slower than FULL (p=.003), while CORNERS did not differ from FULL . So corner-based representations provide additional geometric information without increasing cognitive cost to humans, whereas side-based representations seem to impose greater demands for spatial integration. For both reasoning and non-reasoning models, Table 7 shows the average number of output tokens per representation format. FULL generally produces more output tokens, but higher token usage does not consistently translate into higher accuracy. This asymmetry further differentiates human spatial reasoning from current VLM behavior. Finally, human performance showed no systematic improvement in terms of accuracy and response time over the course of the experiment (Appendix, Fig. 5). Trials were fully randomized, and accuracy did not exhibit a consistent trend as a function of trial index. This observed stability supports the interpretation that the measured human performance reflects single-shot geometric reasoning rather than cumulative practice effects, enabling a clean comparison with model evaluations conducted without training or adaptation.

4.3 ADDITIONAL MODEL ANALYSES OF REPRESENTATION SENSITIVITY

How well do the VLMs do per target position? To assess whether specific target locations (Figure 1a) influence performance, we analyze accuracy per target location aggregated across human and model data (Figure 2).

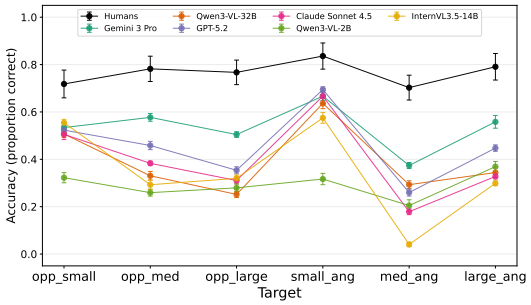

Figure 2: Mean accuracy per target location, aggregated over angle configurations and representations.

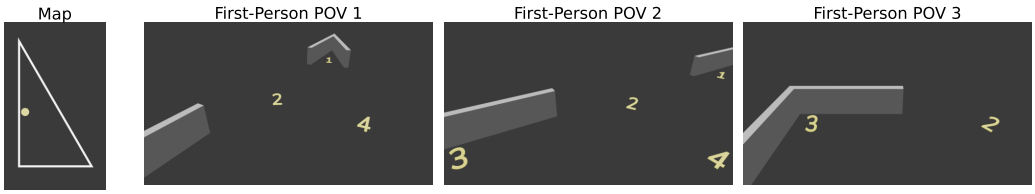

Figure 3: An example of incongruent POV setup: the Full map (same as in Figure 1), paired with Corners first-person views. The correct answer is "2".

Consistent with prior developmental findings (Dillon et al., 2013), performance scales with geometric distinctiveness. Overall, for both humans and models, accuracy is highest when the target corresponds to the smallest angle. More detailed analysis (Appendix, Figure 7) shows that this trend holds in almost all the cases for models, and in the most separable configurations for humans (35-60-85 and 30-60-90).

Although VLMs achieve substantially lower absolute accuracy than humans, they show a highly similar rank ordering of target difficulty. Across most models and configurations, the smallest angle target yields the highest accuracy, while the medium angle target is consistently among the most challenging. This convergence suggests that both humans and current VLMs are sensitive to geometric distinctiveness, even though models do not achieve human-level robustness.

Detailed analyses of accuracy by target, broken down by angle configuration and representation, are shown in Figures 7 and 8, respectively.

How well do the VLMs do when the representation between map and POV views are incongruent? To test whether models rely more on visual consistency between the map (allocentric view) and POV (egocentric view) than on reasoning about the spatial layout depicted in the images, we introduce an *incongruent* setup in which each question preserves the same scene, target, and POV renders but swaps the map representation so that the representation format of the map differs from that of the POV views (e.g., a Corners map paired with a Sides POV). Figure 3 shows an example of this manipulation.

Beyond representation invariance, this setup probes whether a map is actually informative for guiding navigation. Real-world agents must detect when symbolic representations are mismatched and not just rely on superficial visual alignment. This manipulation is inspired by developmental paradigms showing that children flexibly recruit distinct geometric representations when interpreting maps, instead of automatically integrating all cues (Dillon et al., 2013; Dillon & Spelke, 2018). Our incongruent manipulation tests whether VLMs exhibit comparable flexibility or instead depend on superficial representational alignment.

We compared congruent and incongruent settings on 7 models from Table 2 using the same seeds. The results in Table 4 show that incongruence reduced overall accuracy for all models, confirming that no model achieves fully representation-invariant spatial reasoning. The magnitude of the drop,

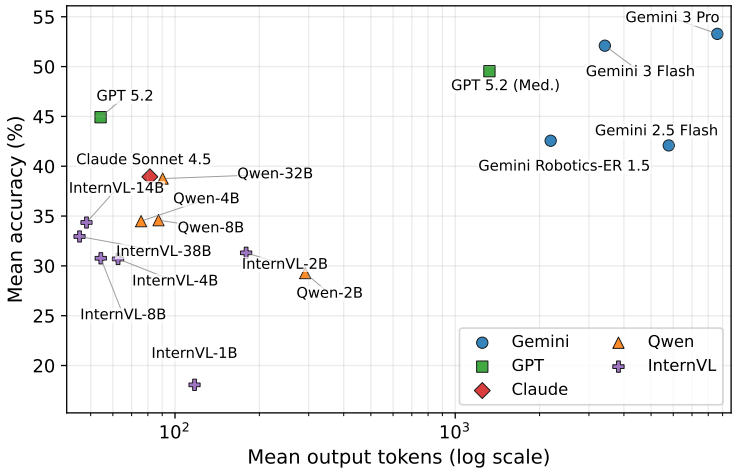

Figure 4: Mean accuracy vs. mean output tokens (log scale) across 18 vision-language models, grouped by model family.

however, was highly model-dependent. The largest drops were in frontier closed-source models: Gemini 3 Pro (-7.6%), Gemini 3 Flash (-6.1%) and GPT 5.2 (-5.9%), while open-source models showed smaller absolute declines but from lower baselines.

The interaction between incongruence and angle configuration is also model-dependent. For some models (e.g., Gemini 3 Pro), drops increase with configuration difficulty, while others (e.g., Gemini 3 Flash) show the opposite pattern, and some open-source models exhibit small local gains in specific configurations despite an overall decrease. Representation-conditioned analyses (map-conditioned vs. POV-conditioned) are reported in Table 6 of the Appendix.

Table 4: Model performance for the incongruent setting across all four angle configurations. For each configuration, we report average accuracy as mean$_{\pm\text{std}}$ over 5 runs (seeds). Arrow annotations show absolute change relative to the congruent setup for the same model and column.

	Avg (A) by angle configuration (%)				
Model	50-60-70	40-60-80	35-60-85	30-60-90	A
Closed-Source Models					
Gemini 3 Pro	47.0$_{\pm1.1}$ (↓1.3)	47.2$_{\pm1.5}$ (↓7.1)	44.4$_{\pm3.5}$ (↓10.4)	44.5$_{\pm1.4}$ (↓11.2)	45.8$_{\pm0.9}$ (↓7.6)
Gemini 3 Flash	45.1$_{\pm1.2}$ (↓8.1)	46.2$_{\pm1.1}$ (↓7.1)	44.7$_{\pm0.8}$ (↓6.1)	48.0$_{\pm1.1}$ (↓3.1)	46.0$_{\pm0.3}$ (↓6.1)
GPT 5.2	38.6$_{\pm1.9}$ (↓4.6)	39.1$_{\pm2.1}$ (↓4.3)	38.3$_{\pm2.5}$ (↓8.7)	40.2$_{\pm1.9}$ (↓5.8)	39.1$_{\pm1.5}$ (↓5.9)
Open-Source Models					
Qwen3-VL-4B	35.7$_{\pm1.5}$ (↑0.7)	32.6$_{\pm2.0}$ (↑0.6)	32.5$_{\pm1.5}$ (↓0.5)	36.9$_{\pm1.2}$ (↓1.1)	34.4$_{\pm1.0}$ (↓0.1)
Qwen3-VL-32B	41.3$_{\pm1.7}$ (↑0.7)	38.8$_{\pm2.6}$ (↑0.3)	35.5$_{\pm1.4}$ (↓2.9)	33.6$_{\pm2.2}$ (↓3.9)	37.2$_{\pm0.8}$ (↓1.5)
InternVL-3.5-14B	29.8$_{\pm1.2}$ (↓1.7)	32.2$_{\pm1.5}$ (↓2.2)	31.3$_{\pm0.9}$ (↓5.0)	32.5$_{\pm1.9}$ (↓2.7)	31.5$_{\pm0.7}$ (↓2.9)
InternVL-3.5-38B	31.2$_{\pm1.6}$ (↓2.9)	32.6$_{\pm2.0}$ (↓2.2)	30.1$_{\pm1.7}$ (↓1.1)	31.7$_{\pm1.9}$ (↔0.0)	31.4$_{\pm1.0}$ (↓1.5)

How does output verbosity relate to accuracy across model families? Figure 4 shows mean accuracy against mean output tokens (in log scale) for 18 VLMs across five families. The highest-performing models, including Gemini 3 Pro ($\sim53\%$), Gemini 3 Flash ($\sim52\%$) and GPT 5.2 with medium reasoning effort ($\sim49\%$), are all reasoning models, occupying the upper-right quadrant.

However, verbosity alone does not guarantee accuracy: Gemini 2.5 Flash produces among the largest number of mean output tokens yet achieves only $\sim42\%$, while GPT 5.2 in standard mode reaches $\sim45\%$ with fewer than 100 tokens, making it the most token-efficient high-performing model on REMAP by a wide margin.

The remaining non-reasoning models, including open-source families (Qwen, InternVL) and Claude Sonnet 4.5, cluster in the low-token regime ($<$300 tokens) with moderate accuracy ($\sim$25%-40%). Within the Qwen family, accuracy scales consistently with model size (2B to 32B), while the trend is noisier for InternVL. Claude Sonnet 4.5 ($\sim$39%) is close to the best open-source models evaluated.

Overall, reasoning models generally achieve the highest accuracy, but the relationship between output length and performance is not monotonic, suggesting that spatial reasoning depends more on model capability and reasoning quality than solely on verbosity.

5 DISCUSSION

Current frontier models, such as Gemini 3 Pro, GPT 5.2, and Claude Sonnet 4.5, achieve high scores on many vision-language tasks, yet they do not reliably and efficiently generalize to robust cross-view geometric reasoning. On REMAP, although Gemini 3 Pro outperforms other models, it still falls behind average human performance and exhibits qualitatively different patterns across representational formats. Moreover, model performance degrades sharply under incongruent map-view pairings, in contrast to human spatial reasoning, which flexibly draws on different geometric cues depending on task demands (Dillon et al., 2013; Dillon & Spelke, 2018). These findings expose a persistent gap in current models' ability to maintain consistent geometric structure when translating between allocentric and egocentric reference frames under varying viewpoints and input formats.

A consistent finding across our experiments is that SIDES representations pose the greatest difficulty for most VLMs, even when angular discriminability is high. This pattern suggests that models may struggle to infer geometric structure from partial edge information, which requires mentally completing corners and integrating segment orientations, a form of constructive spatial inference. Humans, by contrast, perform best with CORNERS representations and adapt their strategy depending on which geometric relations are most informative.

The gap between human and model performance on REMAP has practical implications for embodied agents that must interpret maps, follow spatial instructions, or navigate unfamiliar environments. Our results suggest that current VLMs cannot be reliably deployed for tasks requiring flexible alignment between symbolic spatial descriptions and egocentric visual input. The incongruent analysis further indicates that models may fail silently when map and environment representations do not match, which is a common occurrence in real-world deployment, where maps may be outdated, stylized, or incomplete.

6 CONCLUSION

We introduced REMAP, a benchmark for cross-view geometric reasoning grounded in developmental research on symbolic spatial representation. Evaluating 17 VLMs, we find that leading models outperform random baselines but remain substantially below human accuracy. Beyond an overall performance gap, models exhibit systematic divergences from human behavior: they show persistent weaknesses on side-based representations, limited adaptation to geometric/angular discriminability, and representation formats that do not mirror human sensitivity to geometric structure. These results suggest current VLMs lack robust, geometry-aware mechanisms for aligning allocentric and egocentric representations. While models can partially succeed under certain triangle types and representation formats, their performance does not reflect the structured, adaptive patterns observed in human spatial reasoning.

A key direction for future work is to move beyond static viewpoints toward interactive spatial reasoning. Developmental paradigms, including Dillon & Spelke (2018), often allow participants to navigate and inspect spatial cues. Extending REMAP from static renders generated using 3D setups to interactive 3D settings would enable evaluation of how models (and humans) allocate attention across spatial features, whether active viewpoint selection improves egocentric/allocentric viewpoint alignment, and what spatial cues most effectively support cross-view reasoning. A complementary direction is to embed the same controlled geometric probes within richer, more realistic environments, e.g., with diverse and controlled textures, clutter, and variable lighting. We will release the code and data, which will be updated to reflect ongoing improvements and extended experiments.

ACKNOWLEDGMENTS

We thank the members of the Cornell LIL Lab for their valuable feedback, as well as their advice and support in piloting the human experiments. We thank Susan Epstein for insightful discussions and guidance on human experiment and task design. We are also grateful to friends affiliated with CUNY, Cornell University, UC Berkeley, and beyond for their helpful suggestions and contributions to the human experiments. We thank Google for supporting the experiments with Gemini in this work through a gift. We thank OpenAI for supporting the experiments with GPT in this work through the Researcher Access Program.

REFERENCES

Anthropic. Claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025.

Blender Foundation. Blender. https://www.blender.org, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Judy S DeLoache. Symbolic functioning in very young children: Understanding of pictures and models. *Child development*, 62(4):736–752, 1991.

Judy S DeLoache. Dual representation and young children's use of scale models. *Child development*, 71(2):329–338, 2000.

Moira R Dillon and Elizabeth S Spelke. From map reading to geometric intuitions. *Developmental psychology*, 54(7):1304, 2018.

Moira R Dillon, Yi Huang, and Elizabeth S Spelke. Core foundations of abstract geometry. *Proceedings of the National Academy of Sciences*, 110(35):14191–14195, 2013.

Google DeepMind. Gemini 3. https://deepmind.google/models/gemini/, 2025.

Google DeepMind Robotics. Gemini Robotics ER-1.5, 2025.

Véronique Izard and Elizabeth S Spelke. Development of sensitivity to geometry in visual forms. *Human evolution*, 23(3):213, 2009.

Barbara Landau, Henry Gleitman, and Elizabeth Spelke. Spatial knowledge and geometric representation in a child blind from birth. *Science*, 213(4513):1275–1278, 1981.

Sang Ah Lee, Valeria A Sovrano, and Elizabeth S Spelke. Navigation as a source of geometric knowledge: Young children's use of length, angle, distance, and direction in a reorientation task. *Cognition*, 123(1):144–161, 2012.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *International Conference on Learning Representations (ICLR)*, 2026.

OpenAI. GPT-5.2 System Card. https://openai.com/index/gpt-5-system-card-update-gpt-5-2/, 2025.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024.

Qwen Team. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations (ICLR)*, 2026.

A FULL RESULTS

Table 5 presents the breakdown of the results presented in Table 2.

It highlights three consistent patterns. First, the overall ranking is stable across all four triangle configurations: humans remain well above all VLMs (76.7 average accuracy), Gemini 3 Pro and Gemini 3 Flash are the strongest model families (53.3 and 52.1), and the best open-source model, Qwen3-VL-32B, is still substantially lower at 38.8. Second, representation choice matters almost as much as model choice: for humans and five of the seven closed-source models, *Corners* is the strongest representation, while *Sides* is the weakest for six of the seven closed-source models. Third, the per-representation performance of the open-source models shows more mixed signals but also weaker performance. Overall, a recurring weakness across VLMs is recovering the target from side-based evidence alone, even when performance on Full or Corners is comparatively strong.

Table 6 shows the breakdown of the results presented in Table 4.

Table 5: Full results across all four triangle configurations (50-60-70/40-60-80/35-60-85/30-60-90). For each model and configuration, we report accuracy (%) as mean$_{\pm \text{std}}$ over 5 runs (seeds) for three input representations: Full (F), Corners (C), and Sides (S), plus overall accuracy (A). The rightmost **Avg** block averages each column over the four configurations. Within each configuration block, **green** and light green mark the best and second-best models by overall accuracy (A). Within each model row, **blue** and light blue mark the best and second-best representations (among F/C/S), and pink marks the worst.

	50-60-70				40-60-80				35-60-85				30-60-90				**Avg**			
Model	F	C	S	A	F	C	S	A	F	C	S	A	F	C	S	A	F	C	S	A
Random	29.3	29.3	29.3	29.3	27.9	27.9	27.9	27.9	26.7	26.7	26.7	26.7	27.3	27.3	27.3	27.3	27.8	27.8	27.8	27.8
Human	78.1	75.0	68.4	74.0	65.7	86.4	68.3	74.2	76.1	88.1	85.4	83.0	62.5	83.7	79.6	75.6	70.6	83.3	75.4	76.7
Closed-Source Models																				
Gemini 3 Pro	51.2±3.9	48.8±3.9	44.9±2.8	48.3±2.2	60.5±7.7	52.6±3.5	49.8±3.9	**54.3±3.3**	55.1±3.7	59.6±3.3	49.8±4.9	**54.8±1.7**	58.2±4.4	60.5±3.4	48.6±4.7	**55.8±1.1**	56.2	55.3	48.3	**53.3**
Gemini 3 Flash	48.3±4.0	60.0±2.2	51.2±1.7	**53.2±0.9**	49.3±2.5	59.1±2.1	51.6±3.4	53.3±1.7	49.3±3.7	56.4±1.2	46.7±2.2	50.8±1.1	52.7±4.4	56.8±5.8	43.6±4.1	51.1±3.2	49.9	58.1	48.3	52.1
Gemini 2.5 Flash	46.8±5.6	39.5±5.3	43.9±1.7	43.4±2.6	43.7±5.3	49.3±5.6	38.1±4.5	43.7±3.4	43.1±6.6	51.1±3.5	25.8±6.0	40.0±3.6	43.2±4.8	54.1±4.9	26.4±3.4	41.2±4.2	44.2	48.5	33.5	42.1
Gemini Robotics ER-1.5	38.5±2.0	49.8±5.3	45.9±3.2	44.7±2.0	44.7±1.9	50.7±2.5	35.3±6.2	43.6±2.4	40.9±6.0	52.0±1.2	30.7±4.3	41.2±3.3	42.7±4.1	55.5±4.7	24.1±3.8	40.8±1.6	41.7	52.0	34.0	42.6
GPT-5.2 Medium	47.3±3.7	51.7±3.6	41.0±6.3	46.7±3.2	47.4±5.1	58.1±3.3	41.4±5.0	49.0±3.1	49.8±2.0	59.6±2.9	41.8±2.9	50.4±1.9	50.5±3.0	60.0±3.4	45.9±3.4	52.1±2.3	48.7	57.4	42.5	49.5
GPT-5.2	42.9±4.1	44.9±2.2	42.0±2.7	43.3±2.6	47.0±5.0	49.3±5.6	34.0±2.1	43.4±3.2	48.0±5.6	55.1±3.7	37.8±4.4	47.0±2.6	44.1±4.1	53.6±3.0	40.5±1.9	46.1±2.2	45.5	50.7	38.5	44.9
Claude Sonnet 4.5	44.4±2.0	35.1±2.8	47.8±5.6	42.4±1.6	37.7±4.5	40.9±3.5	37.7±4.2	38.8±1.6	39.1±1.2	37.3±1.9	40.4±2.9	39.0±1.0	35.9±1.0	30.9±3.4	40.0±4.1	35.6±0.9	39.3	36.1	41.5	38.9
Open-Source Models																				
Qwen3-VL-2B	35.1±2.2	24.4±7.1	38.5±5.8	32.7±3.7	24.2±5.4	27.9±9.2	25.1±5.3	25.7±4.6	34.2±2.0	27.6±6.0	28.9±8.2	30.2±2.8	29.5±5.6	26.8±5.9	29.1±3.7	28.5±4.2	30.8	26.7	30.4	29.3
Qwen3-VL-4B	46.8±3.2	27.8±4.4	30.2±2.2	35.0±1.5	38.1±6.5	30.7±1.9	27.0±4.8	31.9±3.5	40.4±5.1	34.7±5.1	24.0±1.9	33.0±2.8	43.2±1.6	38.2±7.4	32.7±4.1	38.0±1.8	42.1	32.8	28.5	34.5
Qwen3-VL-8B	37.6±5.9	38.0±3.7	34.1±7.1	36.6±4.7	40.5±3.1	39.1±3.4	28.8±4.2	36.1±2.3	35.6±8.2	32.4±3.4	27.6±5.8	31.9±3.4	33.2±4.7	36.8±1.9	31.4±3.7	33.8±2.5	36.7	36.6	30.5	34.6
Qwen3-VL-32B	45.4±6.1	42.0±4.7	34.6±4.7	40.7±4.5	42.3±3.0	41.4±1.9	31.6±2.1	38.4±1.4	44.9±4.6	36.9±3.0	33.3±3.1	38.4±1.0	35.5±5.9	45.5±2.3	31.8±3.6	37.6±2.0	42.0	41.4	32.9	38.8
InternVL-3.5-1B	21.0±5.1	18.5±4.1	20.5±2.8	20.0±1.9	17.7±1.3	19.1±6.5	16.7±8.0	17.8±4.6	16.9±2.5	20.0±5.9	18.2±4.3	18.4±2.0	14.5±5.9	16.8±6.7	16.8±6.3	16.1±4.2	17.5	18.6	18.1	18.1
InternVL-3.5-2B	35.1±3.7	26.3±3.2	30.2±5.6	30.6±1.6	28.4±3.0	31.6±5.8	30.7±4.2	30.2±3.1	32.9±2.9	28.4±8.1	31.6±4.0	31.0±3.0	37.7±2.0	30.0±4.9	32.7±4.4	33.5±1.5	33.5	29.1	31.3	31.3
InternVL-3.5-4B	31.2±4.4	26.8±4.6	30.2±2.8	29.4±3.7	34.9±5.9	27.9±4.9	29.8±6.9	30.9±4.0	32.0±4.6	30.2±3.7	27.1±4.8	29.8±1.6	32.7±6.1	33.6±1.9	31.8±3.2	32.7±1.5	32.7	29.6	29.7	30.7
InternVL-3.5-8B	34.1±1.7	25.4±3.7	30.2±5.1	29.9±2.4	37.7±1.9	30.2±6.8	28.8±4.8	32.2±2.8	33.3±4.2	25.8±1.2	24.4±4.2	27.9±2.1	35.5±2.6	34.5±3.7	29.1±4.4	33.0±1.7	35.2	29.0	28.2	30.8
InternVL-3.5-14B	27.3±2.0	30.7±2.8	36.6±3.9	31.5±0.9	35.8±4.5	27.9±1.6	39.5±2.3	34.4±2.1	36.9±1.2	30.7±1.9	41.3±4.6	36.3±1.4	36.8±3.0	33.2±5.0	35.5±1.2	35.2±1.6	34.2	30.6	38.2	34.4
InternVL-3.5-38B	35.6±3.3	27.3±3.2	39.5±5.8	34.1±1.4	42.3±1.0	34.4±3.0	27.4±3.8	34.7±1.0	38.3±4.9	29.4±6.1	26.1±4.2	31.3±2.8	32.7±4.1	34.1±3.6	28.2±2.6	31.7±2.4	37.2	31.3	30.3	33.0

Table 6: Incongruent setup results with per representation breakdowns. Left block reports mean$_{\pm \text{std}}$ over 5 runs (seeds) by angle configuration. Right blocks report mean accuracy grouped by map representation and by POV representation, respectively. **green** and light green mark best and second-best model per setup/overall column; within each model and within each representation block, **blue** and light blue mark best and second-best representation, and pink marks worst (ties share rank).

	Avg (A) by angle configuration (%)					Avg by rep. (Map)			Avg by rep. (POV)		
Model	50-60-70	40-60-80	35-60-85	30-60-90	A	F	C	S	F	C	S
Closed-Source Models											
GPT-5.2	38.6±1.9	39.1±2.1	38.3±2.5	40.2±1.9	39.1±1.5	38.3	32.7	46.3	44.2	41.4	31.6
Gemini 3 Pro	47.0±1.1	**47.2±1.5**	44.4±3.5	44.5±1.4	45.8±0.9	54.9	40.6	41.7	50.6	41.6	45.1
Gemini 3 Flash	45.1±1.2	46.2±1.1	**44.7±0.8**	**48.0±1.1**	**46.0±0.3**	52.0	42.2	43.9	51.5	43.8	42.7
Open-Source Models											
Qwen3-VL-4B	35.7±1.5	32.6±2.0	32.5±1.5	36.9±1.2	34.4±1.0	36.0	33.4	33.9	36.9	33.6	32.7
Qwen3-VL-32B	41.3±1.7	38.8±2.6	35.5±1.4	33.6±2.2	37.2±0.8	37.6	35.1	38.8	40.5	38.3	32.8
InternVL-3.5-14B	29.8±1.2	32.2±1.5	31.3±0.9	32.5±1.9	31.5±0.7	31.9	31.7	30.9	31.0	28.9	34.6
InternVL-3.5-38B	31.2±1.6	32.6±2.0	30.1±1.7	31.7±1.9	31.4±1.0	32.5	30.6	31.1	31.8	30.5	31.8

A.1 Additional Results on Human Experiments

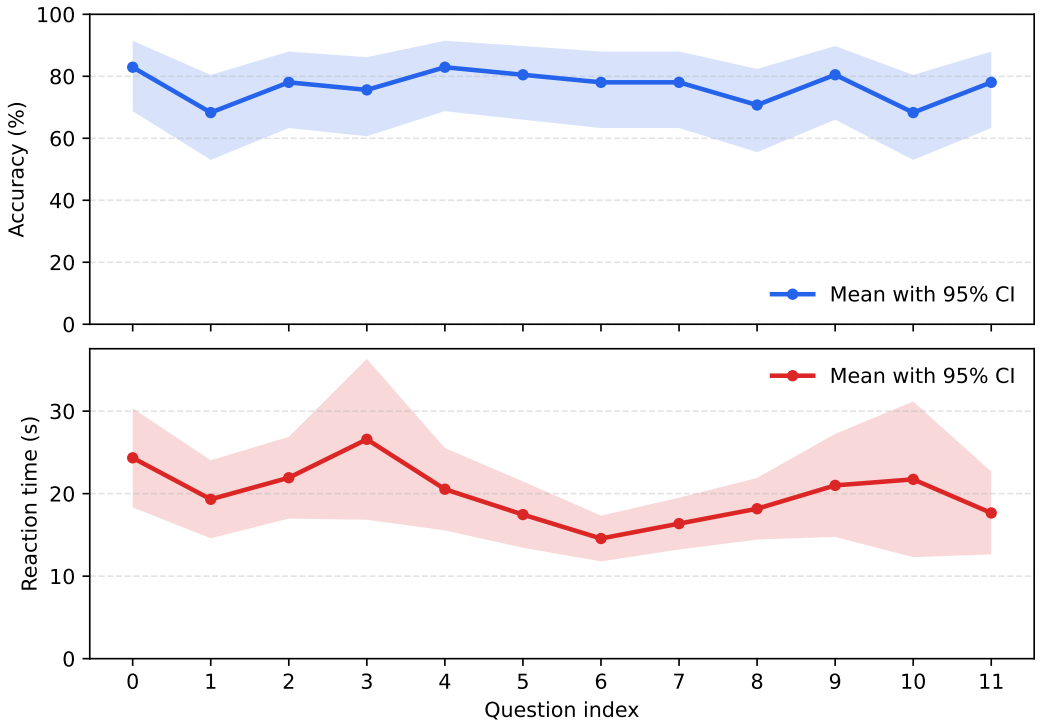

Figure 5: Human performance over time. Top panel shows mean accuracy by question index. Bottom panel shows mean reaction time (s) by question index.

B Human Evaluation and Data Collection

B.1 Prompts

Below we provide the task instructions shown to participants.

Instructions to the human participants

You are given a map showing an overhead view of a scene.

In the scene, a yellow circle marks a target location.

You'll also see three images that represent what you would see from a first-person view (as if you were standing in the scene and looking from three slightly different angles).

In these images, each number represents a unique location in the scene. This means that if the same number appears in multiple first-person view images, it refers to the same physical location seen from different angles. For example, if you see the number "8" in two out of the three images, both instances represent the same location labeled "8", just viewed from two different angles.

One and only one of the numbers matches the location marked on the map by the yellow circle.

The numbers are different in each question.

Your task: Choose the numbered location in the images that corresponds to the yellow circle's location on the map.

B.2 Interface

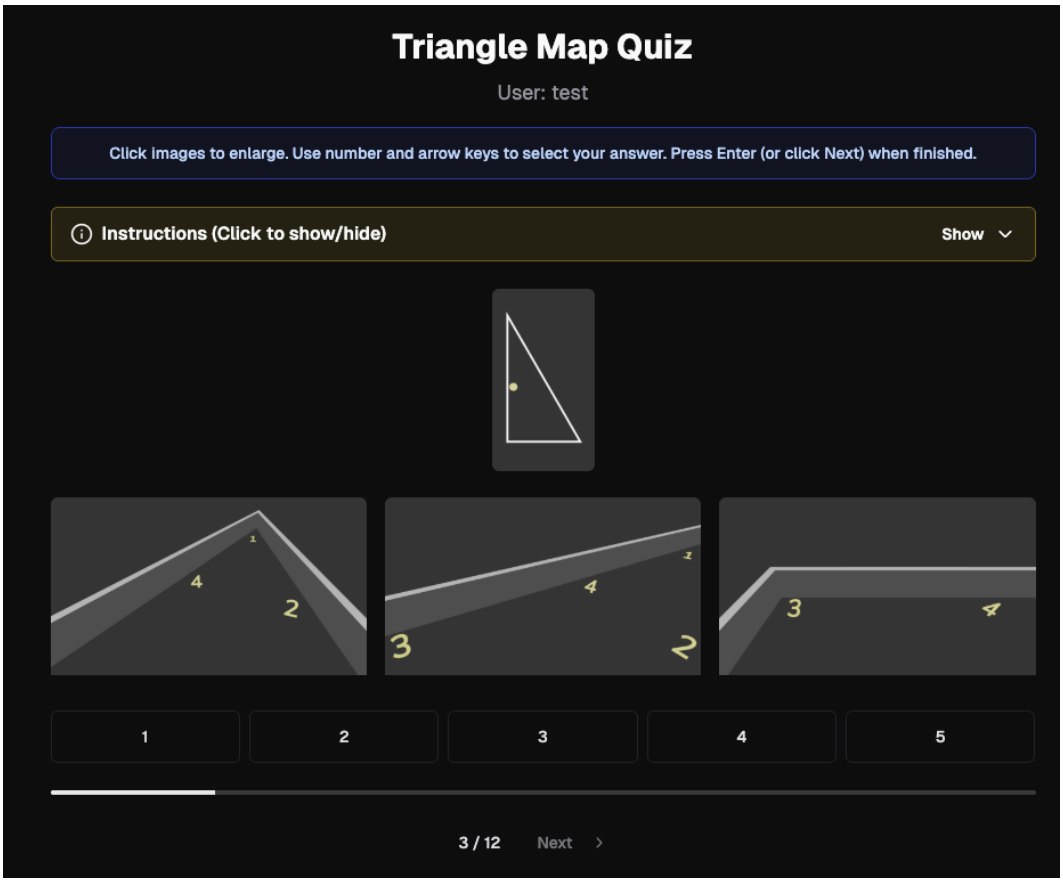

Figure 6: Screenshot of the interface provided to the human participants.

C Average Accuracy and Output Tokens

Table 7: Average accuracy and output tokens by representation for selected models.

Model	Representation	Avg accuracy (%)	Avg output tokens ± std
Gemini 3 Pro	Full	56.30	9516.21 ± 126.77
Gemini 3 Pro	Corners	55.49	8497.99 ± 460.44
Gemini 3 Pro	Sides	48.32	7872.86 ± 282.63
Gemini 3 Flash	Full	49.94	3702.56 ± 599.69
Gemini 3 Flash	Corners	58.03	3177.95 ± 253.22
Gemini 3 Flash	Sides	48.21	3354.05 ± 243.14
GPT 5.2 Medium	Full	48.79	1483.47 ± 87.05
GPT 5.2 Medium	Corners	57.46	1125.77 ± 180.28
GPT 5.2 Medium	Sides	42.54	1360.17 ± 157.51
GPT 5.2	Full	45.55	54.53 ± 1.05
GPT 5.2	Corners	50.87	54.25 ± 0.94
GPT 5.2	Sides	38.50	54.02 ± 0.64
Qwen3-VL-32B	Full	41.97	95.18 ± 3.66
Qwen3-VL-32B	Corners	41.39	81.33 ± 1.44
Qwen3-VL-32B	Sides	32.83	94.48 ± 3.72

D Aggregated Accuracy per Target

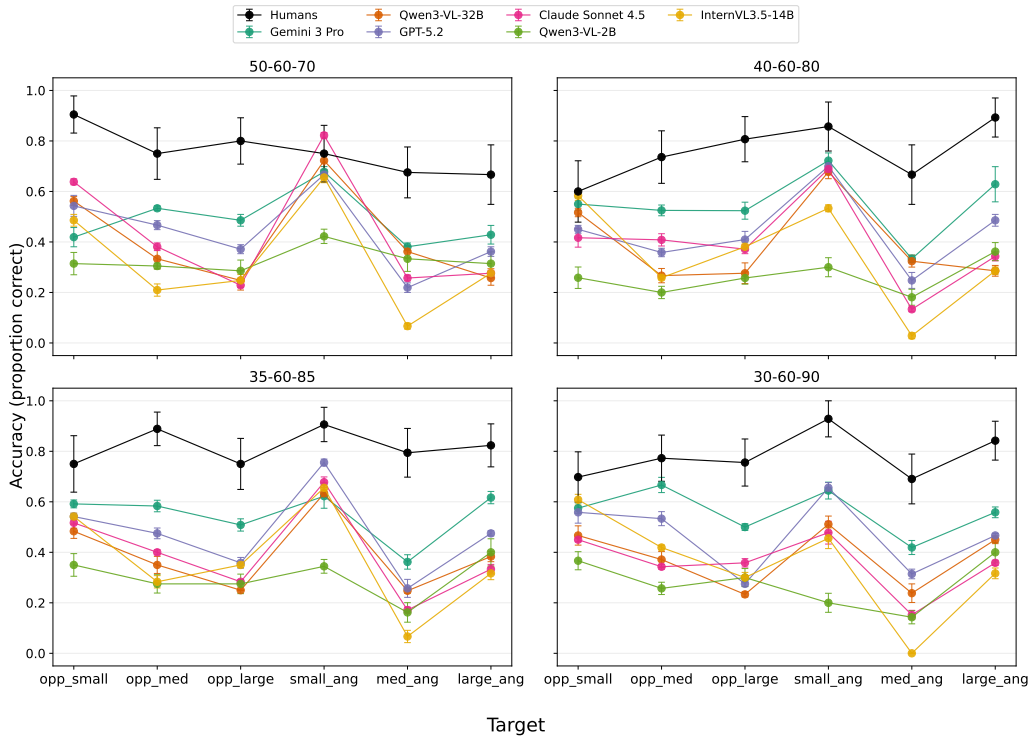

Figure 7: Aggregated mean accuracy per target location by angle configuration

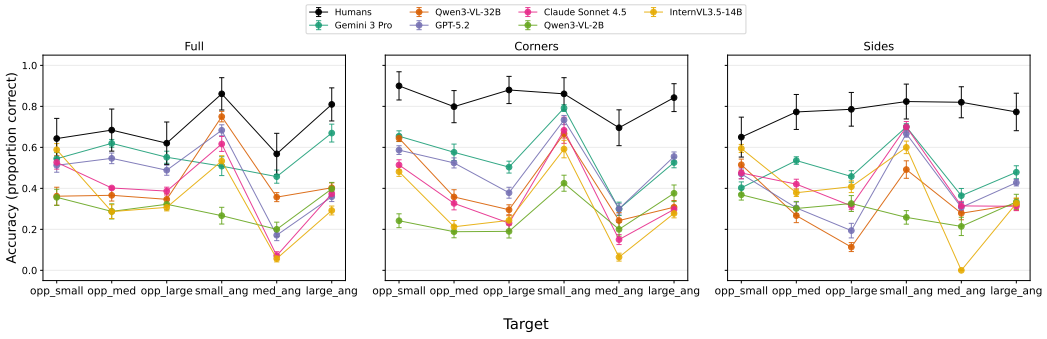

Figure 8: Aggregated mean accuracy per target location by representation format