REMAP: EVALUATING CROSS-PERSPECTIVE SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Building coherent spatial representations requires relating what is seen from different viewpoints to a shared global layout. Frontier multimodal models are becoming increasingly capable at this problem in realistic environments, but it is unclear whether these gains reflect robust spatial reasoning or dependence on familiar correspondences between representations. We introduce REMAP, a benchmark for cross-perspective spatial reasoning in which models identify a target location on an allocentric map from multiple egocentric views. REMAP is based on settings developed for cognitive studies of human spatial reasoning. It combines FLOORPLAN, which is based on indoor layouts, and TRIANGLE, a controlled diagnostic that isolates robustness to changes in how local views correspond to the global map. REMAP exhibits high discriminative power. Across 24 vision-language models and human participants, capability ranges from below random to surpassing average human performance on FLOORPLAN. TRIANGLE enables deeper analysis, showing that simple variations in the geometric correspondence between views can lead to significant performance drops, whereas humans are largely robust.

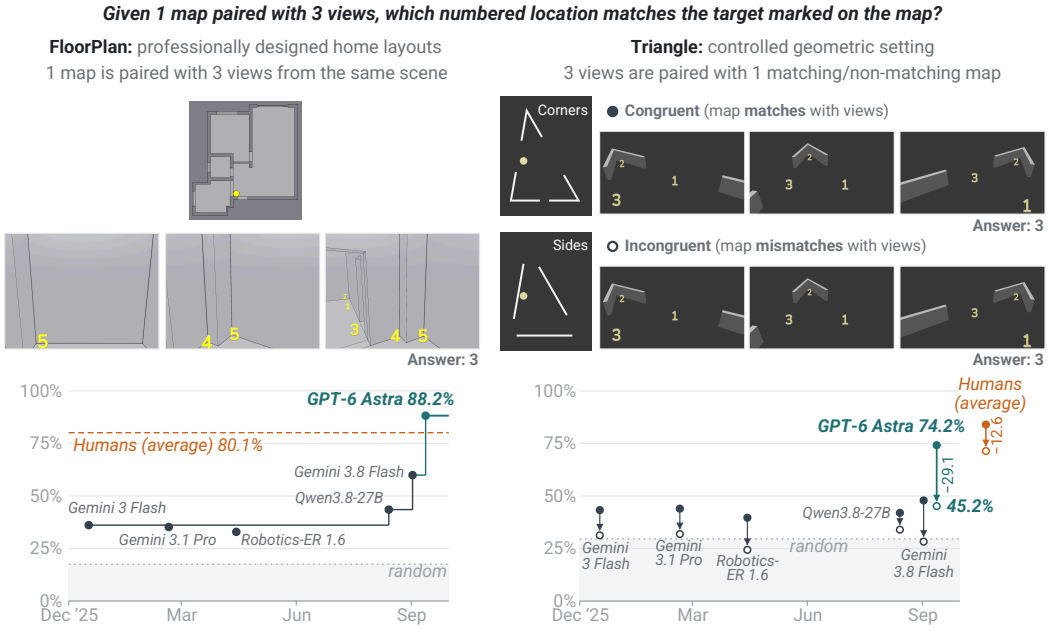

Figure 1: Overview of REMAP. **Top**: models match locations across allocentric maps and egocentric views in realistic indoor layouts (FloorPlan) and controlled geometric scenes (Triangle). **Bottom:** comparison of models and human performance on REMAP, showing a step change in cross-perspective spatial reasoning in frontier models (left). Yet changing how the same geometry is depicted in Triangle produces a much larger performance drop for the strongest models (right).

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

1 INTRODUCTION

Spatial correspondence requires identifying the same locations across viewpoints and representations, even when their visual appearances differ substantially. People handle this reliably in routine problems: they relate each separated part in an exploded assembly diagram to its place in the assembled object, and match the schematic strokes of a hand-drawn sketch to the scenes they observe. For example, a room sketch may mark only the middle sections of its walls, leaving the corners undrawn, and the reader must infer that these gaps reflect the drawing convention rather than openings in the room. Successful correspondence requires determining which differences arise from the underlying scene and which arise from how it is depicted. Although Vision-Language Models (VLMs) have advanced considerably in spatial reasoning, high accuracy on richly detailed scenes may depend on familiar representations and semantic cues such as object identity and texture (Chen et al., 2024; Li et al., 2026b; Ramakrishnan et al., 2025). Such performance alone does not establish how reliably models relate the same underlying geometry across different depictions.

To study these questions, we introduce REMAP, a cross-perspective spatial reasoning benchmark that forces the models to rely on geometric reasoning, with graded levels of layout realism and pairs every setting with human performance on the same questions. The visual scenes are stripped of object and texture cues, leaving layout geometry as the basis for correspondence. REMAP includes two settings. FLOORPLAN presents the task on professionally designed, realistic indoor layouts, while TRIANGLE is a controlled, cognitive science-based diagnostic setting, targeting the use of distance and angle as distinct foundations of geometric reasoning (Dillon et al., 2013; Dillon & Spelke, 2018).

Across extensive experiments on 24 closed and open VLMs, we find that strong spatial performance coexists with substantial sensitivity to these changes. REMAP demonstrates strong discriminative ability, showing that on realistic FLOORPLAN, GPT-6 Astra achieves human-level performance. Yet on the minimal TRIANGLE setting, a change in representation yields a drop in performance of 29.1%, while humans maintain a difference of 12.6%. Human participants show substantially greater success on these mismatches without being explicitly taught the convention.

We observe that targeted guidance can help recover lost performance in strongest models, showing strong perceptual ability. However, we argue that beyond final accuracy, evaluating a model's capabilities requires considering how much information and assistance it needs to complete a task. In the case of humans, similar or better accuracy is achieved without further guidance. Error analyses further show that guidance can reduce confusion between target families while leaving errors in identifying the correct location.

Our contributions are threefold: (1) REMAP, a benchmark combining realistic map-to-view correspondence, answer-preserving changes in representation, and human performance on corresponding questions; (2) evidence that strong spatial accuracy can coexist with substantial sensitivity to representation and dependence on interpretive guidance; and (3) controlled instruction and error analyses that characterize what this guidance resolves and which difficulties remain. Together, these results motivate evaluating spatial capabilities through both the outcomes models achieve and the information and assistance under which they achieve them.

2 RELATED WORK

Spatial reasoning and maps in VLMs. Broad benchmarks show that spatial reasoning remains a weakness of current VLMs, from perceptual correspondence and perspective taking to visual simulation (Fu et al., 2024b; Ramakrishnan et al., 2025; Stogiannidis et al., 2025; Li et al., 2026b; Jia et al., 2026). Map benchmarks such as MAPWise, MapEval, CartoMapQA, FRIEDA and MapQA test how well models read cartographic maps, including symbols, legends, scale and routes (Mukhopadhyay et al., 2025; Dihan et al., 2025; Ung et al., 2025; Pyo et al., 2026; Arnold et al., 2026). Success there can depend on map conventions and place knowledge rather than on relating a map to what is seen on the ground. Closer to our task, m2sv asks models to infer the camera heading by aligning a north-up map with a Street View image (Shin et al., 2026). Multi-view benchmarks, including VSI-Bench, MindCube and Theory of Space, ask whether models can build a spatial representation from video, a few views or active exploration (Yang et al., 2025; Gholami et al., 2025; Wang et al., 2026; Zhang et al., 2026), and SCOPE tests whether spatial judgments stay consistent across perspectives and viewpoints (Kim et al., 2026). Separately, IsoBench shows that

accuracy on the same problems changes with how they are represented, with models consistently more accurate on text than on images (Fu et al., 2024a). REMAP brings this question of representation to map-to-view correspondence: matched TRIANGLE questions differ only in how the map is drawn, while the views, candidate labels and correct answer stay fixed.

Cognitive foundations of spatial reasoning. Using a picture or scale model to stand for a larger space requires treating it as a representation of that space and not only as an object in its own right, which young children find difficult (DeLoache, 1991; 2000). Using purely geometric maps, Dillon et al. (2013) found that children's navigation predicted their map use when targets were defined by distances, whereas their analysis of visual forms predicted map use when targets were defined by corner angles, with no evidence of an integrated representation of distance and angle. Dillon & Spelke (2018) then showed that six-year-olds struggle to use a map depicting only a triangle's sides in an environment made of only its corners, and vice versa; this limitation weakens by about age ten.

3 REMAP: THE REPRESENTATIONAL MAPPING BENCHMARK

To test whether a model keeps a spatial correspondence when the same space is drawn differently, REMAP poses one map-to-view task in two settings. FLOORPLAN places the task in realistic multi-room homes and measures how well models solve it in everyday layouts. TRIANGLE places it in a controlled enclosure whose walls can be redrawn, so that the map can change while everything else in the question stays the same (Figure 2).

3.1 THE TASK

Figure 1 shows one question from each setting. Each question has two parts. The first is a top-down map of the scene, on which a yellow circle marks the target. The second is a set of three first-person views taken inside the same scene with visible candidate locations represented by integer labels. The task is to find which integer label in the three views corresponds to the target location.

Two rules make the answer well defined. First, when the same physical location appears in several views, it is consistently represented by the same integer label. Second, the target is always visible in at least one view. The answer is derived from the scene geometry, so it requires no manual annotation. The scenes contain no furniture or texture, hence the correspondence has to rely on the layout itself rather than on recognizable objects. Because the number of visible candidates differs between questions, we compare accuracy with an item-aware random baseline (Section 4).

3.2 FLOORPLAN: THE REALISTIC INDOOR SETTING

FLOORPLAN asks how well models solve the task in layouts like those people use every day. Its 291 homes are reconstructed from Structured3D annotations and rendered in Blender (Zheng et al., 2020; Blender Foundation, 2025). They keep walls, doorways, windows and the connections between rooms, but not furniture or texture. The map shows the whole home, while the three views come from cameras inside a single room. A model must therefore match what it sees locally to the plan of the whole home, even when walls and doorways hide parts of it. Targets lie at room corners and at the midpoints of walls, and map and views always draw the walls the same way. The setting contains 3,603 questions with 2 to 23 candidates, grouped into four candidate-count bins (Appendix A).

3.3 TRIANGLE: THE CONTROLLED SETTING

TRIANGLE asks a narrower question: does a model that finds the target with one drawing of a space still find it when the map is drawn differently? To answer it, the scene must be simple enough that the drawing can be changed without changing anything else. TRIANGLE uses the triangular enclosure of developmental studies of geometric map reading (Dillon et al., 2013; Dillon & Spelke, 2018). It has four scalene triangles, each with a 60° angle (Appendix A), and six target roles in each: the three corners and the midpoints of the three sides. As in FLOORPLAN, the enclosure has no furniture or texture.

Each triangle is built in three wall formats, which make different parts of the shape visible:

- **FULL (F):** continuous walls, so both corners and sides are visible.
- **CORNERS (C):** short walls at each corner, with gaps along the sides.
- **SIDES (S):** one wall segment centred on each side, with the corners left open.

The wall format does not depend on the target role, so a corner target can appear with any wall format. REMAP's Congruent and Incongruent setup adapts the map-environment matching manipulation of Dillon & Spelke (2018), who compared maps and enclosures depicting the same or complementary features of a triangle, i.e., its corners or sides.

Redrawing only the map. Figure 2 (top) walks through one example. A camera triplet renders three CORNERS views of a fixed target and numbers the visible candidates (step 2). The same three views, with the same labels, are then paired with three maps of the same triangle: a CORNERS map, a FULL map and a SIDES map (step 3). With the CORNERS map, map and views are drawn alike, and we call the question **Congruent**. With the FULL or SIDES map, the maps are drawn differently from the views, and we call it **Incongruent**. Taken together, the three questions show the same scene, the same target and the same answer (1). They differ only in how the map is drawn. A model that finds the target in the Congruent question but not in the Incongruent ones therefore fails because of the drawing alone.

All map and view pairings. The example holds the views in the CORNERS format. Doing the same for FULL and SIDES views, and pairing each of the three map formats with each of the three view formats, gives nine combinations (Figure 2, step 4): three Congruent and six Incongruent. Only combinations that share the same views isolate the effect of the map. Changing the view format produces new views and can renumber the candidates.

Size. We extend this design to include FULL depictions of both maps and views. Each of the 24 triangle and target pairs is rendered from 18 camera triplets, which vary how far apart the cameras are. Each of these 432 renderings appears in all nine combinations, giving 3,888 questions (1,296 Congruent and 2,592 Incongruent). Table 1 covers all nine combinations, and some later analyses use only the CORNERS and SIDES formats. The appendix gives the full construction and examples of every combination (Appendix A).

The two settings play complementary roles. FLOORPLAN measures performance in realistic layouts, and TRIANGLE isolates the effect of redrawing the map. Because the settings also differ in size and in the number of candidates, we do not attribute differences between them to realism alone.

4 EXPERIMENTAL SETUP

Models and human study. We report on 24 models across 5 API systems (GPT 6 Astra, Gemini 3.8 Flash, Gemini 3.1 Pro, Gemini Robotics-ER 1.6, Gemini 3 Flash) and 19 open models, spanning Qwen, InternVL, Gemma, and Molmo and including the spatial specialists SpatialEvo-7B and ViLaSR (Google DeepMind, 2025; 2026c;a;d; OpenAI, 2026; Bai et al., 2025; Qwen Team, 2025; 2026; Wang et al., 2025; Google DeepMind, 2026b; Clark et al., 2026; Li et al., 2026a; Wu et al., 2025). For human evaluation, we randomly sampled a subset of questions from each benchmark. The TRIANGLE subset includes 576 questions, and the FLOORPLAN subset includes 600 questions. Each question was answered by three English-speaking adult participants from Prolific.

Metrics. We report raw answer accuracy and a random baseline. For a question i with c_i candidate labels, uniform guessing succeeds with probability $1/c_i$, so the baseline is $\frac{1}{N}\sum_i 1/c_i$. Raw TRIANGLE and FLOORPLAN accuracies should therefore be read against their own baselines. We evaluate all models with 3 runs across different seeds (except for GPT-6 Astra, with only one run) using each model's recommended sampling configuration. Responses without a valid location label are counted as incorrect. We report the mean and standard deviation when relevant.

Prompts. We iteratively refined the model instructions to clarify the multi-image task, the identity of repeated location labels across views, and the requirement to select a single answer. We then fixed one prompt for all models and both settings, requesting a one-line explanation followed by a location label. Human instructions conveyed the same core task information, adapted to the study interface.

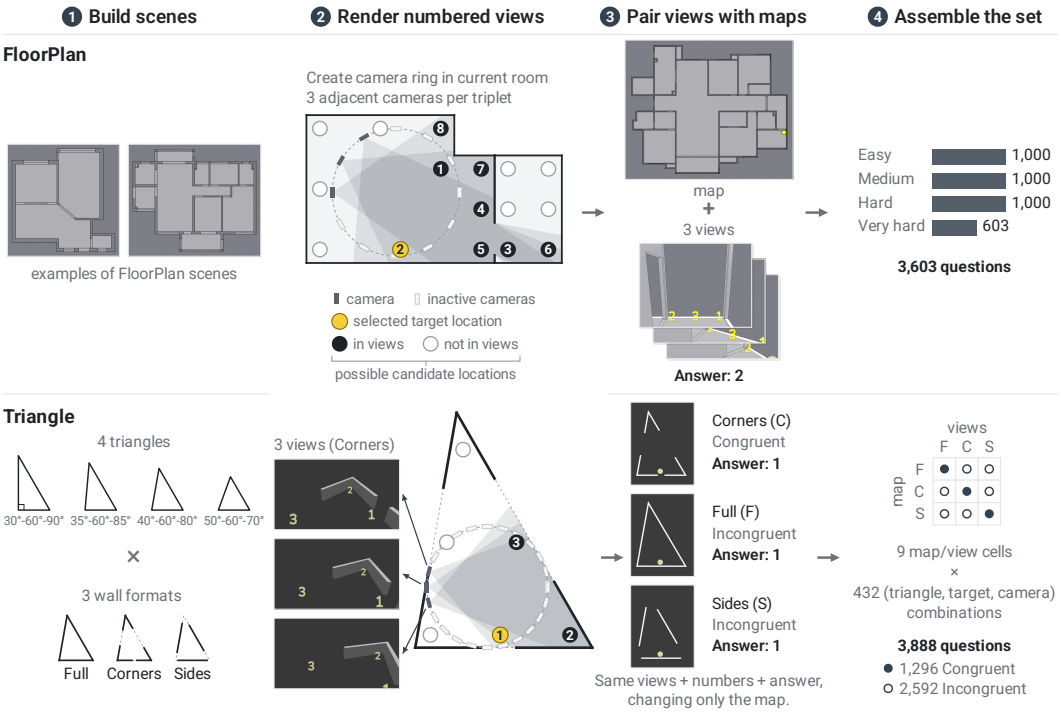

Figure 2: How REMAP questions are generated for FLOORPLAN (*top*) and TRIANGLE (*bottom*). **(1) Build scenes.** Both settings use Blender. FLOORPLAN reconstructs 291 professionally designed home layouts. TRIANGLE creates four triangles whose angles grow progressively harder to tell apart, from 30-60-90 to 50-60-70, each drawn with FULL (F), CORNERS (C), or SIDES (S) walls. **(2) Render numbered views.** Candidate locations are defined at each corner and the midpoint of each side. Three adjacent cameras form a triplet of views, and only visible candidate locations are numbered. **(3) Pair views with maps.** Each triplet of views is paired with a map marking a target location. In TRIANGLE, the same views, e.g., CORNERS, can be paired with CORNERS, FULL, and SIDES maps, varying only the map while keeping the views, numbers, and ground truth constant. **(4) Assemble the set.** FLOORPLAN contributes 3,603 questions across four candidate-count difficulty bins. Pairing 3 map formats with 3 view formats gives 9 cells of 432 blocks, or 3,888 TRIANGLE questions (1,296 Congruent, 2,592 Incongruent).

Crucially, neither the human instructions nor the baseline model prompt explained how the Triangle drawing formats correspond. Full instructions are provided in Appendices B.2 and C.

5 RESULTS

5.1 STRONG PERFORMANCE ON REALISTIC LAYOUTS DOES NOT CARRY ACROSS REPRESENTATIONS

Table 1 reports model accuracy on the full benchmarks and human accuracy on the study subsets. At a high level, the competence that the strongest models show on realistic layouts of FLOORPLAN neither carries over to the minimal TRIANGLE scenes nor survives a change in how the map is drawn, whereas human competence largely does, when provided with the same amount of information about the task. The pattern is sharpest for GPT-6 Astra, the strongest model: it surpasses the average human response on FLOORPLAN (88.2% vs. 80.1%), falls behind on Congruent TRIANGLE questions (74.2% vs. 84.0%), and shows the largest Incongruent penalty of any system (29.1 points), more than twice the average human penalty (12.6 points).

Table 1: Accuracy (%) on the full REMAP benchmark for both the FLOORPLAN and TRIANGLE settings for every evaluated model. TRIANGLE Congruent/Incongruent: map and views drawn the same way or differently (all pairings of FULL, CORNERS and SIDES). Δ: accuracy lost when the map is redrawn (Congruent−Incongruent), and darker red means a larger drop. Grey numbers are below the random baseline.

		FLOORPLAN	TRIANGLE		
Model / reference	Released		Congruent	Incongruent	Δ
● GPT-6 Astra	Sep 2026	**88.2**	**74.2**	**45.2**	29.1
● Gemini 3.8 Flash	Sep 2026	59.9±0.4	47.9±0.7	28.3±0.2	19.6±0.7
● Gemini Robotics-ER 1.6	Apr 2026	32.9±0.4	39.7±0.6	24.4±0.2	15.2±0.6
● Gemini 3.1 Pro	Feb 2026	35.2±0.9	44.0±0.9	31.9±0.9	12.1±0.9
● Gemini 3 Flash	Dec 2025	36.1±0.8	43.3±0.3	31.3±0.2	12.0±0.1
○ Qwen3.8-27B	Aug 2026	43.5±0.4	42.0±0.8	34.0±1.3	8.0±0.7
○ Qwen3-VL-32B	Oct 2025	18.7±0.1	28.9±0.5	28.1±0.2	0.8±0.6
○ Qwen3-VL-4B	Oct 2025	18.6±0.6	23.6±1.2	23.4±0.2	0.2±1.1
○ Qwen3-VL-8B	Oct 2025	18.6±0.3	27.3±1.4	26.2±1.0	1.1±2.4
○ Qwen2.5-VL-32B	Mar 2025	17.6±0.3	23.4±0.4	22.6±0.7	0.8±1.1
○ Qwen2.5-VL-3B	Jan 2025	17.1±0.3	27.9±1.0	27.8±0.6	0.1±1.3
○ Qwen2.5-VL-7B	Jan 2025	17.4±0.2	24.5±1.8	23.6±0.4	0.8±1.4
○ InternVL2.5-38B	Dec 2024	18.8±0.3	28.7±1.0	28.8±1.3	-0.1±0.4
○ InternVL2.5-8B	Dec 2024	17.2±0.4	25.8±0.9	25.0±0.3	0.8±1.2
○ InternVL3-38B	Apr 2025	18.6±0.8	31.7±0.1	29.4±0.7	2.2±0.6
○ InternVL3-8B	Apr 2025	17.4±0.3	26.2±1.2	24.7±0.5	1.6±1.6
○ InternVL3.5-38B	Aug 2025	19.4±0.2	26.2±1.0	22.9±0.3	3.3±0.7
○ InternVL3.5-8B	Aug 2025	17.7±0.2	22.3±0.7	22.6±0.6	-0.3±0.5
○ Gemma 4 26B-A4B-it	Apr 2026	21.0±0.1	29.3±0.8	25.5±0.1	3.8±0.8
○ Gemma 4 E2B-it	Apr 2026	17.3±0.2	24.7±0.3	24.8±0.4	-0.1±0.3
○ Gemma 4 E4B-it	Apr 2026	17.0±0.2	23.2±0.1	22.7±0.2	0.5±0.3
○ Molmo2-8B	Dec 2025	17.0±0.3	26.4±2.5	25.1±1.0	1.3±2.1
○ SpatialEvo-7B	Apr 2026	16.6±0.5	24.0±1.4	22.9±1.2	1.1±1.8
○ ViLaSR	Jun 2025	17.4±0.5	23.6±0.8	21.8±0.2	1.8±0.6
● Humans (average)		80.1	84.0	71.4	12.6
Random baseline		17.5	29.6	29.6	0.0

The first gap appears between settings. The strongest models are much better on realistic layouts than on minimal ones, whereas humans are not. Measured relative to chance,[1] GPT-6 Astra and Gemini 3.8 Flash score 22–25 points higher on FLOORPLAN than on Congruent TRIANGLE questions, Qwen3.8-27B 14 points, and the earlier Gemini models at most 4, while the average human response scores about the same on both. Accuracy on realistic layouts thus overstates general spatial competence, and increasingly so for stronger models.

Within TRIANGLE, a second and larger gap appears when the map is drawn differently from the views. Here every frontier model loses most of its margin above chance: GPT-6 Astra and Qwen3.8-27B lose 65% of their Congruent margin, Gemini 3 Flash and 3.1 Pro over 80%, and Gemini 3.8 Flash and Robotics-ER 1.6 all of it, falling below chance. The average human response loses 23%. Conversely, the near-zero penalties of the other open models do not indicate robustness: with Congruent accuracy near chance, they have no margin to lose.

Newer and more specialized models do not close this gap: from Gemini 3 Flash to Gemini 3.8 Flash, FLOORPLAN accuracy rises by 24 points while Incongruent accuracy falls below chance (31.3% to 28.3%), and Gemini Robotics-ER 1.6 and the spatially trained open models also remain at or below chance on Incongruent questions.

[1]Normalized accuracy is $(a - c)/(100 - c)$, where a is accuracy and c the random baseline (17.5% on FLOORPLAN, 29.6% on TRIANGLE). The share of the margin lost under a redraw is $(a_{\text{Con}} - a_{\text{Inc}})/(a_{\text{Con}} - c)$.

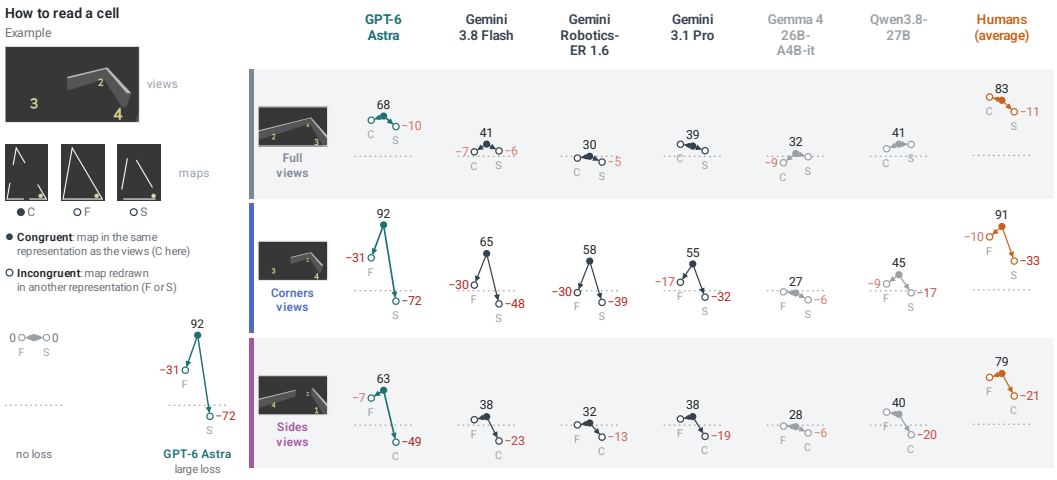

Figure 3: Accuracy change in the Incongruent condition of TRIANGLE when only the map is redrawn (models: full benchmark; humans: 576-question subset). Rows give the view representation and columns the systems. Filled dots are Congruent accuracy, and arrows point to the two Incongruent cells with the same views (map F, C or S). Labels give Incongruent minus Congruent accuracy in points. Dotted lines mark chance: 29.6% for models (Astra one run, others three-run means) and 30.5% for humans.

5.2 REDRAWING ONLY THE MAP CAN SHARPLY REDUCE ACCURACY

Every Incongruent question has a Congruent twin with the same views, target, candidates and answer; only the map is redrawn. Figure 3 shows all six such directions. Because the views are held fixed, each paired change isolates the effect of the map's depiction.

Redrawing the map does more than make questions harder: it steers models to wrong answers on questions they otherwise solve. Redrawing a CORNERS map as SIDES costs Astra 72.5 points (92.1% to 19.7%) and the reverse redraw 49.3 points (62.7% to 13.4%), and every plotted model ends below chance in both directions, almost entirely by overturning answers it had solved.

More surprisingly, a more complete map does not help. Replacing a CORNERS map with the FULL outline, which contains those corners, still lowers Astra from 92.1% to 60.9%, while the average human response moves from 90.6% to 80.7%. This refines the gap identified in Section 5.1: most of the loss arises when a CORNERS map is redrawn as SIDES or the reverse, rather than from any mismatch between map and views. Since the views and answer never change, we next ask how much of this loss targeted guidance can recover.

5.3 TASK-SPECIFIC GUIDANCE RECOVERS WHAT HUMANS ACHIEVE WITHOUT IT

Accuracy alone does not reveal how much information and assistance a model needs to succeed. In practical settings, such as following an assembly diagram, the relevant convention is rarely stated and must be worked out as part of the larger task. Because we constructed the convention of TRIANGLE, we can instead supply it in graded amounts and measure how much of the loss returns.

We add this information, which human participants never received, to the task prompt and evaluate Astra, Gemini 3.8 Flash, Gemini 3 Flash and Qwen3.8-27B on the 576 human-study questions of TRIANGLE (Figure 4).

Targeted information recovers much of Astra's loss: once told what to look for, it answers most redrawn questions correctly. The neutral paragraph changes little, but pointing to the drawings raises Astra's Incongruent accuracy from 55.5% under Neutral to 75.0% under Attention, and stating the convention raises it to 79.9%, while its Congruent accuracy stays between 83.9% and 86.5%. The Congruent–Incongruent gap thus falls from 31.0 points under Neutral to 3.9 points under Explicit convention. The same assistance helps other models far less. Under Explicit convention,

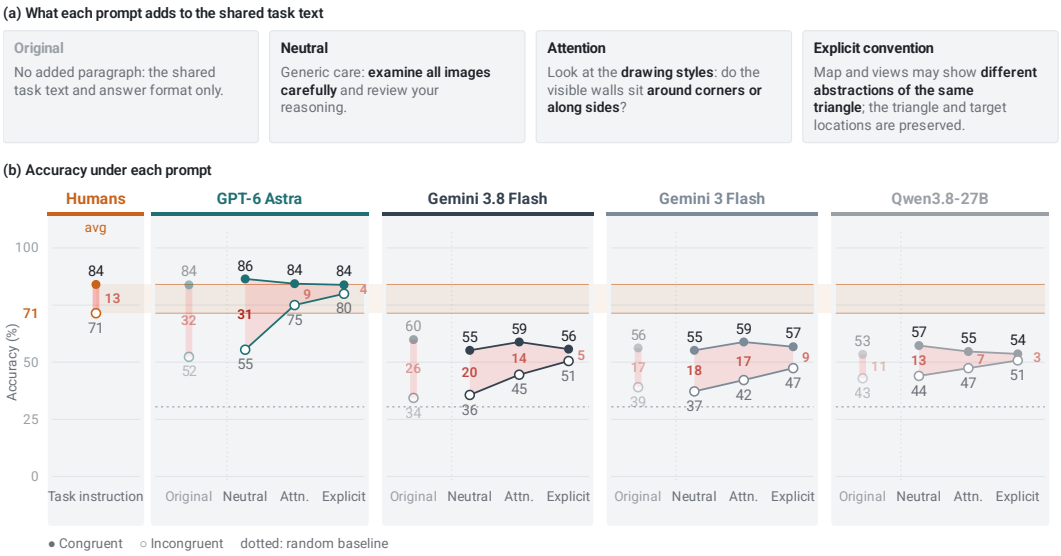

Figure 4: Each prompt adds additional task-specific guidance. *Neutral* asks for careful comparison and controls for the added length, *Attention* points to whether walls appear around corners or along sides, *Explicit convention* states that map and views may depict the same triangle in different ways. Panel (a) paraphrases each added paragraph.

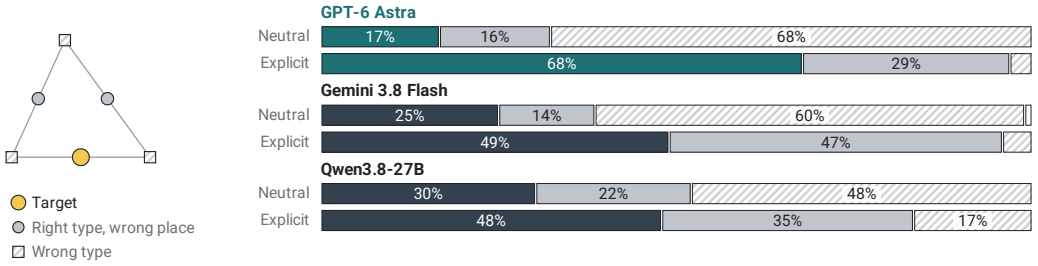

Figure 5: Error composition on the Corners/Sides Incongruent human-study questions that offer both kinds of wrong answer, under Neutral and Explicit convention. Filled: correct. Grey: right type, wrong place (another side midpoint when the target is a side midpoint, or another corner when it is a corner). Hatched: wrong type (a corner chosen for a side midpoint, or the reverse). Left: schematic with the target at a side midpoint.

Gemini 3.8 Flash, Gemini 3 Flash and Qwen3.8-27B reach only 47.4–50.8% on Incongruent questions, against 79.9% for Astra. Recovery therefore separates models more sharply than failure does: all fall below chance on the hardest redraws, but only Astra approaches its Congruent accuracy once guided.

Human participants, by contrast, worked out the convention themselves. Neither they nor the models were told how the drawings relate, yet the average human response is correct on 71.4% of Incongruent questions, a level Astra reaches only once the Attention paragraph points it to the drawings.

Guidance mainly repairs how models read the map, not how they locate the target on it. A wrong answer either picks the wrong type of location (a corner for a side midpoint, or the reverse) or the right type in the wrong place. Figure 5 shows that most of the improvement comes from choosing the right type, but only Astra consistently converts the corrected type into the correct location.

5.4 Human-model comparison reveals shared difficulty but different errors

On the subset of 600 FloorPlan questions human participants answered, Astra answers 89.0% correctly, above the human mean of 80.1%. Its errors nonetheless track human difficulty. Grouping

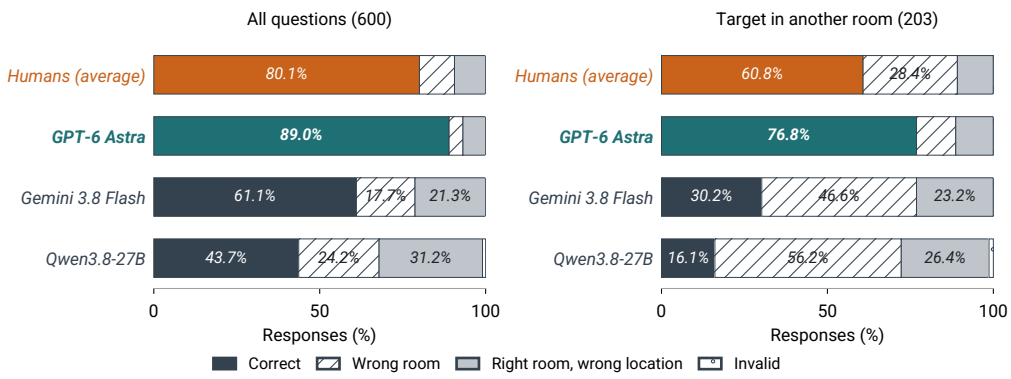

Figure 6: Response outcomes on FLOORPLAN for all 600 questions (left) and the 203 questions whose target lies outside the camera's room (right). Bars partition all responses into correct, wrong-room, correct-room wrong-location, and invalid answers. Humans: three responses per question; Astra: one run; Gemini 3.8 Flash and Qwen3.8-27B: means over three runs. All models use the Original prompt.

items by how many of the three human responses are correct, Astra's error rate rises monotonically across the four groups, from 16 of 369 items with three correct responses (4.3%) to 11 of 28 items with none (39.3%). Higher accuracy therefore does not remove sensitivity to the items humans find difficult.

On the 203 items whose target lies outside the camera's room, Astra gives far fewer wrong-room answers than humans (11.8% vs. 28.4%), while its correct-room, wrong-location rate is descriptively similar on these 203 items (11.3% vs. 10.8%).

On the same 203 questions, Gemini 3.8 Flash and Qwen3.8-27B show a different profile: both exceed human rates for wrong-room answers (46.6% and 56.2%) and for correct-room, wrong-location answers (23.2% and 26.4%). Astra's advantage over humans is thus concentrated in fewer wrong-room answers, whereas the weaker models make more errors of both types. Overall accuracy and error profile therefore provide complementary characterizations of performance.

6 DISCUSSION AND CONCLUSION

In this work, we introduced REMAP, showing that strong map-to-view correspondence on realistic layouts can coexist with sensitivity to answer-preserving redraws, whereas humans provide a more robust behavioral reference.

On the 576-question diagnostic subset, targeted guidance is an effective mitigation for Astra, recovering most of its loss on redrawn maps, whereas other models benefit substantially less. This result separates solving the task once the relevant convention is supplied from establishing correspondence without such help. Human participants received no explicit explanation of the convention yet remained comparatively accurate on redrawn maps. The distinction matters in use, since a person marking a location on a rough sketch need not know which aspect of the depiction needs explaining. The assistance a system needs is therefore part of what an evaluation measures.

High overall accuracy and answers of the right target type do not by themselves establish reliable spatial correspondence. Evaluating that capability also requires testing consistency across depictions and dependence on interpretive help. More broadly, these findings motivate evaluating whether models can identify and apply the assumptions a task depends on, rather than only whether they succeed once those assumptions are supplied.

Ethics statement

The human study recruited participants through Prolific. Participants gave explicit, informed consent before the task. The human-subjects component was reviewed and determined to be exempt. Compensation was calibrated to a base rate of $12/hour of expected completion time, with a correctness bonus of up to $0.80.

Reproducibility statement

We specify the task, the design axes, the generation procedure, per-condition item counts, chance baselines and a construction audit (Appendix A). All the materials, including the data, prompts, and evaluation code will be released publicly upon acceptance.

AI use statement

We used generative AI tools to polish the writing, to debug the code that generates the synthetic benchmark data, and to debug and inspect the evaluation and analysis code. They were not used to design the benchmark or its questions. All ground-truth answers are derived deterministically from scene geometry at generation time. The authors are responsible for validating the final claims, references, and implementations.

References

Christian Michael Arnold, Andrew Alini, Jonathan Wang, Pieter M. Feenstra, Conner Arnold, Jan DeWitt, Natalie C. Ritsema, Jung Hyun Yae, Boris Katz, Andrei Barbu, and Brian Cheung. MapQA: A map-question-answering benchmark for visual language model reasoning. In *ICLR 2026 Workshop on Multimodal Intelligence*, 2026. URL https://openreview.net/forum?id=EEKJXEJJKK.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian 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.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. *arXiv preprint arXiv:2601.10611*, 2026.

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.

Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. In *International Conference on Machine Learning*, pp. 13774–13813, 2025.

Moira R Dillon and Elizabeth S Spelke. From map reading to geometric intuitions. *Developmental psychology*, 54(7):1304–1316, 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.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking multimodal foundation models on isomorphic representations. In *First Conference on Language Modeling (COLM)*, 2024a.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024b.

Mohsen Gholami, Ahmad Rezaei, Weimin Zhou, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with Vision-Language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

Google DeepMind. Gemini 3 Flash. https://deepmind.google/models/model-cards/gemini-3-flash/, 2025.

Google DeepMind. Gemini 3.1 Pro, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemma 4. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/, 2026b.

Google DeepMind. Gemini 3.8 Flash, September 2026c. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Google DeepMind. Gemini Robotics-ER 1.6, 2026d. URL https://deepmind.google/models/model-cards/gemini-robotics-er-1-6/.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for Vision Language models. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Dinging Li, Yingxiu Zhao, Xinrui Cheng, Kangheng Lin, Hongbo Peng, Hongxing Li, Zixuan Wang, Yuhong Dai, Haodong Li, Jia Wang, et al. SpatialEvo: Self-evolving spatial intelligence via deterministic geometric environments. *arXiv preprint arXiv:2604.14144*, 2026a.

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*, 2026b.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. Mapwise: Evaluating Vision-Language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 9348–9378, 2025.

OpenAI. GPT-6 Astra Model. https://developers.openai.com/api/docs/models/gpt-6-astra, 2026. Accessed September 26, 2026.

Jiyoon Pyo, Yuankun Jiao, Dongwon Jung, Zekun Li, Leeje Jang, Sofia Kirsanova, Jina Kim, Yijun Lin, Qin Liu, Junyi Xie, Hadi Askari, Nan Xu, Muhao Chen, and Yao-Yi Chiang. Frieda: Benchmarking multi-step cartographic reasoning in Vision-Language models. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, 2025.

Yosub Shin, Michael Buriek, and Igor Molybog. m2sv: A scalable benchmark for map-to-street-view spatial reasoning. arXiv:2601.19099, 2026.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in Vision-Language models. *arXiv preprint arXiv:2503.19707*, 2025.

Huy Quang Ung, Guillaume Habault, Yasutaka Nishimura, Hao Niu, Roberto Legaspi, Tomoki Oya, Ryoichi Kojima, Masato Taya, Chihiro Ono, Atsunori Minamikawa, and Yan Liu. CartoMapQA: A fundamental benchmark dataset evaluating Vision-Language models on cartographic map understanding. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 440–453, 2025.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Fei-Fei Li, and Manling Li. MindCube: Spatial mental modeling from limited views. In *International Conference on Learning Representations*, 2026.

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.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in Vision-Language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems*, 2025.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 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, Fei-Fei Li, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations (ICLR)*, 2026.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large Photo-Realistic dataset for structured 3D modeling. In *European Conference on Computer Vision*, pp. 519–535. Springer, 2020.

A DATASET CONSTRUCTION, STATISTICS AND EXAMPLES

This appendix gives the size and composition of both settings and shows example questions. Table 2 gives the totals.

A.1 QUESTION FORMAT AND CHANCE LEVEL

Each question has four images: a top-down map followed by three first-person views. The map marks the target with a yellow circle, and the views show numbered candidate locations. A number that appears in several views names one location, so it counts once. The answer is the number of the candidate at the target. The number of candidates K differs between questions, so the chance level is the average probability of guessing each question correctly:

$$\text{Chance} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i}. \tag{1}$$

Table 2: Number of questions in REMAP. TRIANGLE questions are Congruent when map and views are drawn alike and Incongruent otherwise. FLOORPLAN always draws map and views alike. The counts are questions, not model requests or human responses.

Setting	Scenes	Congruent	Incongruent	Total
TRIANGLE	triangle with angles 50-60-70°	324	648	972
	triangle with angles 40-60-80°	324	648	972
	triangle with angles 35-60-85°	324	648	972
	triangle with angles 30-60-90°	324	648	972
	all four triangles	1,296	2,592	3,888
FLOORPLAN	291 homes	not varied		3,603
Both settings				7,491

A.2 TRIANGLE

Triangles, targets and wall formats. The four triangles have angles of 50-60-70, 40-60-80, 35-60-85 and 30-60-90 degrees. Each triangle has six targets: its three corners and the midpoints of its three sides. Each triangle is built in three wall formats. FULL (F) draws all three walls. CORNERS (C) keeps two short wall segments at each corner and leaves gaps along the sides. SIDES (S) keeps one segment centred on each side and leaves the corners open. The target positions are the same in all three formats, so a corner target can appear with any wall format.

Combining map and view formats. The map and the three views each use one of the three formats, which gives nine combinations (Table 3). We write the map format first, so CS means a CORNERS map with SIDES views. When map and views use the same format, the question is Congruent. Otherwise it is Incongruent. Each triangle and target is rendered from 18 camera triplets, which gives 432 sets of views. Each set appears in all nine combinations, so every combination has 432 questions, 108 per triangle and 216 per target type. When only the map changes, the question keeps the same view images, candidate numbers and answer.

Table 3: TRIANGLE questions for every combination of map format and view format. Each combination has 108 questions per triangle and 216 per target type. The target type (corner or side midpoint) is separate from the wall format. The last three rows add up the nine rows above.

			Triangle angles (°)				Target		
Map	Views	Type	50-60-70	40-60-80	35-60-85	30-60-90	Corner	Side	Total
Full	Full	Congruent	108	108	108	108	216	216	432
Full	Corners	Incongruent	108	108	108	108	216	216	432
Full	Sides	Incongruent	108	108	108	108	216	216	432
Corners	Full	Incongruent	108	108	108	108	216	216	432
Corners	Corners	Congruent	108	108	108	108	216	216	432
Corners	Sides	Incongruent	108	108	108	108	216	216	432
Sides	Full	Incongruent	108	108	108	108	216	216	432
Sides	Corners	Incongruent	108	108	108	108	216	216	432
Sides	Sides	Congruent	108	108	108	108	216	216	432
Congruent subtotal			324	324	324	324	648	648	1,296
Incongruent subtotal			648	648	648	648	1,296	1,296	2,592
All nine combinations			972	972	972	972	1,944	1,944	3,888

Candidates. A candidate is one of the six target positions that is visible in at least one of the three views. Questions have between 2 and 6 candidates (Table 4). The candidates depend only on the views, so all nine combinations have the same distribution and the same chance level of 29.6%.

Table 4: TRIANGLE questions by number of candidate locations. A location seen in several views counts once. All nine map and view combinations have the same distribution, because they share the same camera views. Chance is the average of 1 divided by the number of candidates.

	Number of candidates						
Questions	2	3	4	5	6	Total	Chance (%)
Each combination	17	210	169	32	4	432	29.59
All nine combinations	153	1,890	1,521	288	36	3,888	29.59

Examples of all nine combinations. Figures 7 to 15 show one complete set of nine questions, all built from the same 40-60-80 triangle, target and camera triplet. Each page keeps the view format fixed and shows the Congruent question followed by its two Incongruent versions. Within a page, only the map changes and the answer stays the same. Across pages the same target carries a different number, because each view format has its own numbering. The images are shown unchanged, and they were chosen without looking at model results.

Examples with Full POVs held fixed

FF: Full map / Full POVs (Congruent)

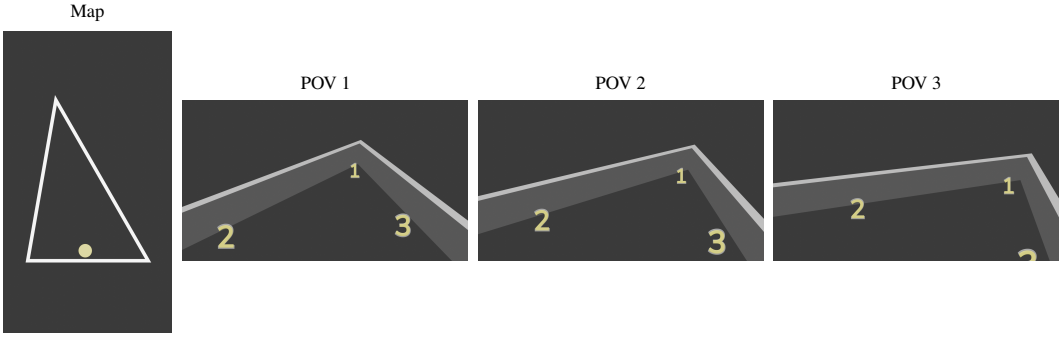

Figure 7: FF example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **3**. This condition contains 432 questions, 108 per shape.

CF: Corners map / Full POVs (Incongruent)

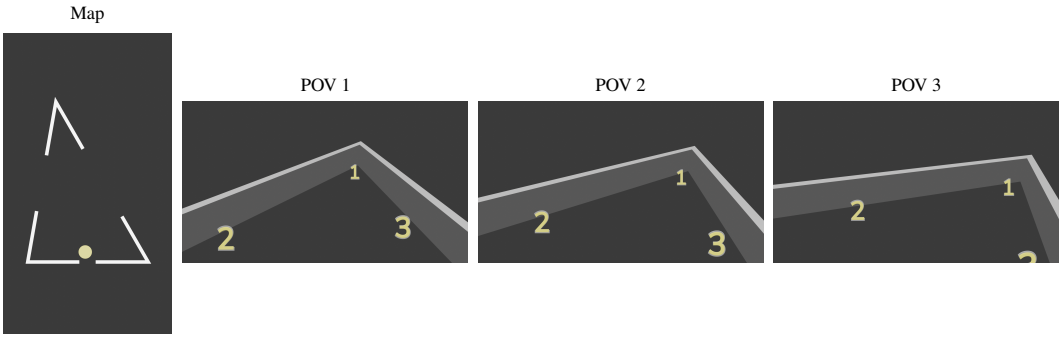

Figure 8: CF example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **3**. This condition contains 432 questions, 108 per shape.

SF: Sides map / Full POVs (Incongruent)

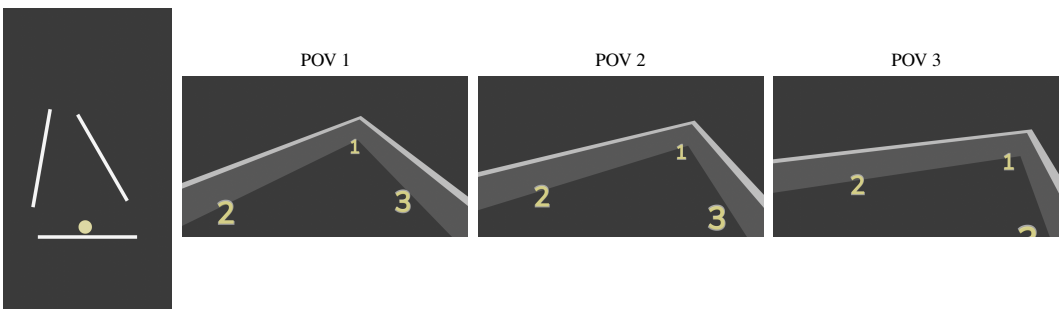

Figure 9: SF example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **3**. This condition contains 432 questions, 108 per shape.

Examples with Corners POVs held fixed

CC: Corners map / Corners POVs (Congruent)

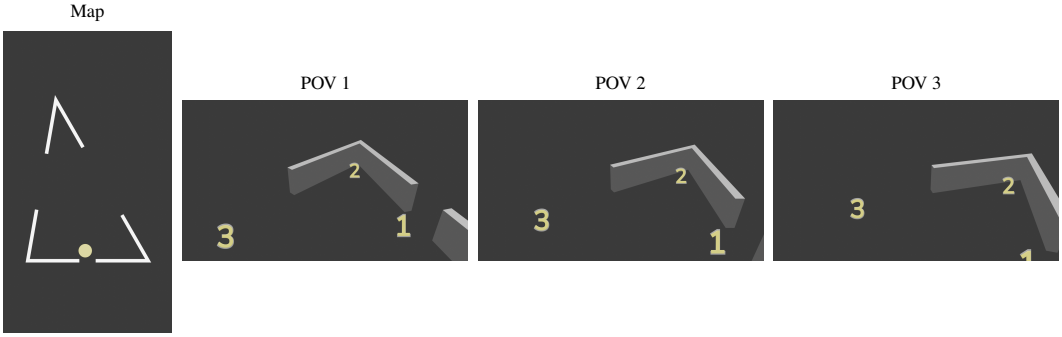

Figure 10: CC example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **1**. This condition contains 432 questions, 108 per shape.

FC: Full map / Corners POVs (Incongruent)

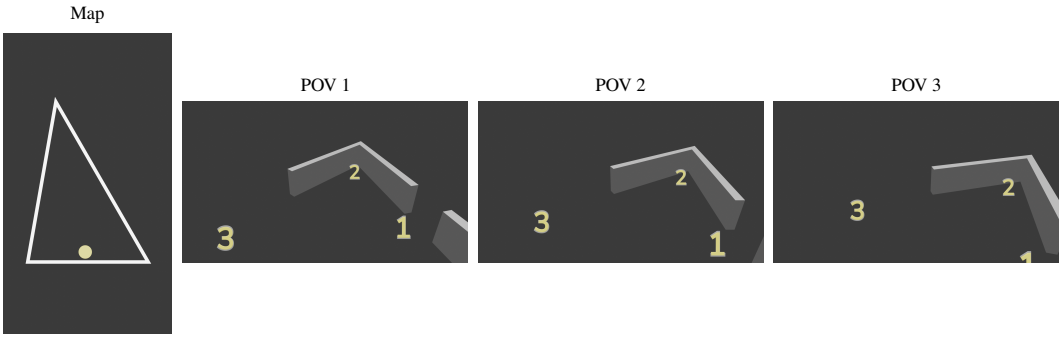

Figure 11: FC example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **1**. This condition contains 432 questions, 108 per shape.

SC: Sides map / Corners POVs (Incongruent)

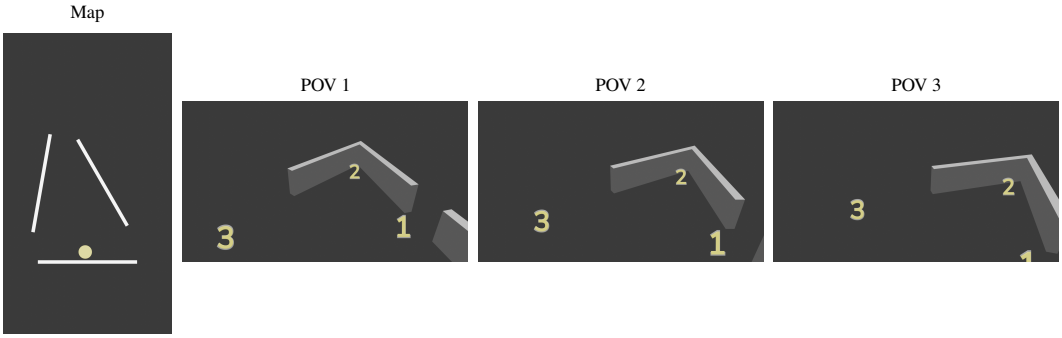

Figure 12: SC example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **1**. This condition contains 432 questions, 108 per shape.

Examples with Sides POVs held fixed

SS: Sides map / Sides POVs (Congruent)

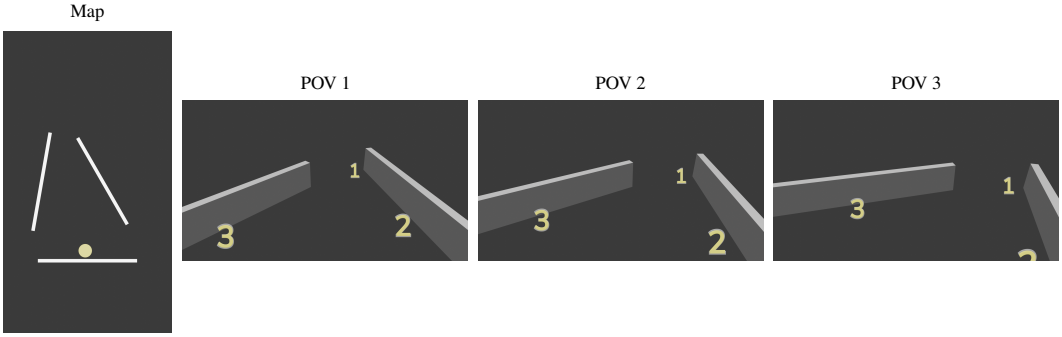

Figure 13: SS example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **2**. This condition contains 432 questions, 108 per shape.

FS: Full map / Sides POVs (Incongruent)

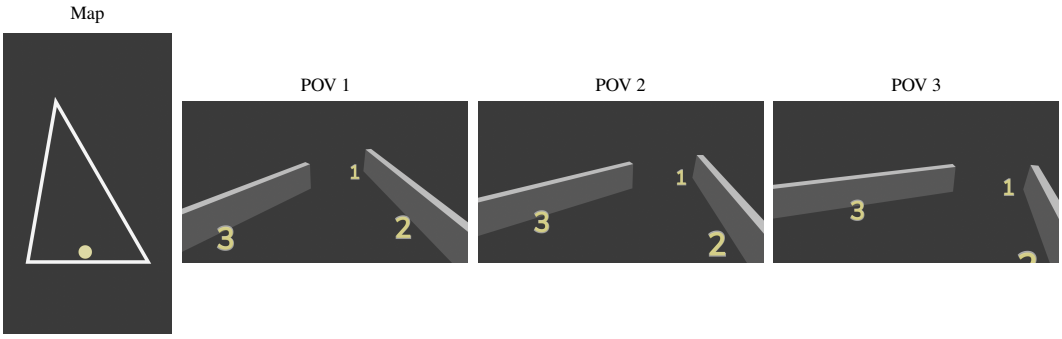

Figure 14: FS example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **2**. This condition contains 432 questions, 108 per shape.

CS: Corners map / Sides POVs (Incongruent)

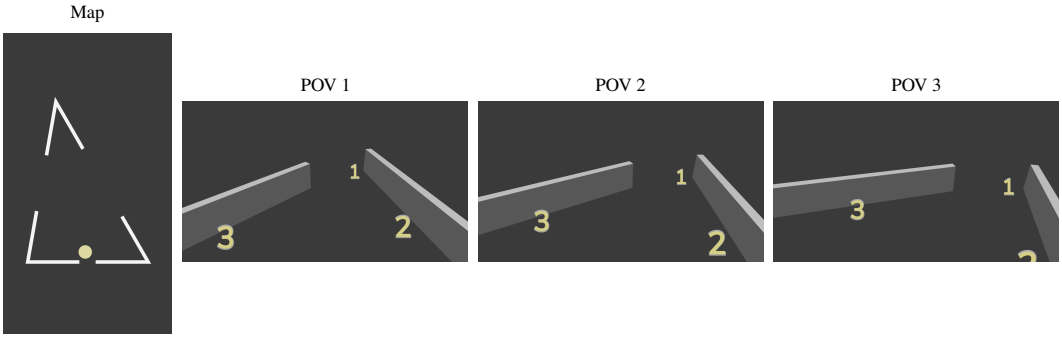

Figure 15: CS example. The target is opposite the $40°$ angle of a $(40, 60, 80)°$ triangle. The three visible candidate locations are numbered 1–3; the correct answer is **2**. This condition contains 432 questions, 108 per shape.

A.3 FLOORPLAN

Homes. We reconstruct the walls, openings and room connections of homes from Structured3D annotations (Zheng et al., 2020) and render them in Blender (Blender Foundation, 2025), without furniture or texture. We sampled 300 homes, stratified by number of rooms. Nine of them produced no questions, which leaves 291 (Table 5).

Table 5: FLOORPLAN homes by number of rooms. Nine of the 300 sampled homes produced no questions, which leaves 291.

Rooms per home	Homes sampled	Homes used	Questions
1-3	45	45	532
4-6	90	90	1,091
7-10	90	90	1,116
11-15	45	38	477
16+	30	28	387
Total	300	291	3,603

Candidates and difficulty. Targets lie at room corners and at the midpoints of room walls. The three views come from neighbouring positions on a ring of 12 cameras inside one room. A candidate is a target position visible in at least one of the three views, and it carries the same number in every view. We group questions into four difficulty levels by number of candidates: Easy (2 to 4), Medium (5 to 8), Hard (9 to 12) and Very hard (13 or more). Each level aims for 500 corner and 500 side targets, with at most one question per room and camera triplet. The first three levels reach 1,000 questions each. Very hard reaches 603, because we do not reuse camera triplets to fill it (Table 6). Over all 3,603 questions, the chance level is 17.5%.

Table 6: FLOORPLAN questions by number of candidate locations. Each difficulty level has 500 corner and 500 side targets, except Very hard (13 or more candidates), which has 603 questions in total. Chance is the average of 1 divided by the number of candidates.

		Target			
Difficulty	Candidates	Corner	Side	Total	Chance (%)
Easy	2-4	500	500	1,000	32.02
Medium	5-8	500	500	1,000	16.73
Hard	9-12	500	500	1,000	10.09
Very hard	13-23	302	301	603	6.83
Total	2-23	1,802	1,801	3,603	17.47

Examples. Figures 16 to 19 show one example from each difficulty level, alternating corner and side targets. Each is the first selected question of its level and target type, chosen without looking at model results. They have 3, 6, 12 and 13 candidates. The images are shown unchanged.

FLOORPLAN EXAMPLES: EASY AND MEDIUM

Easy: 3 candidates, corner target

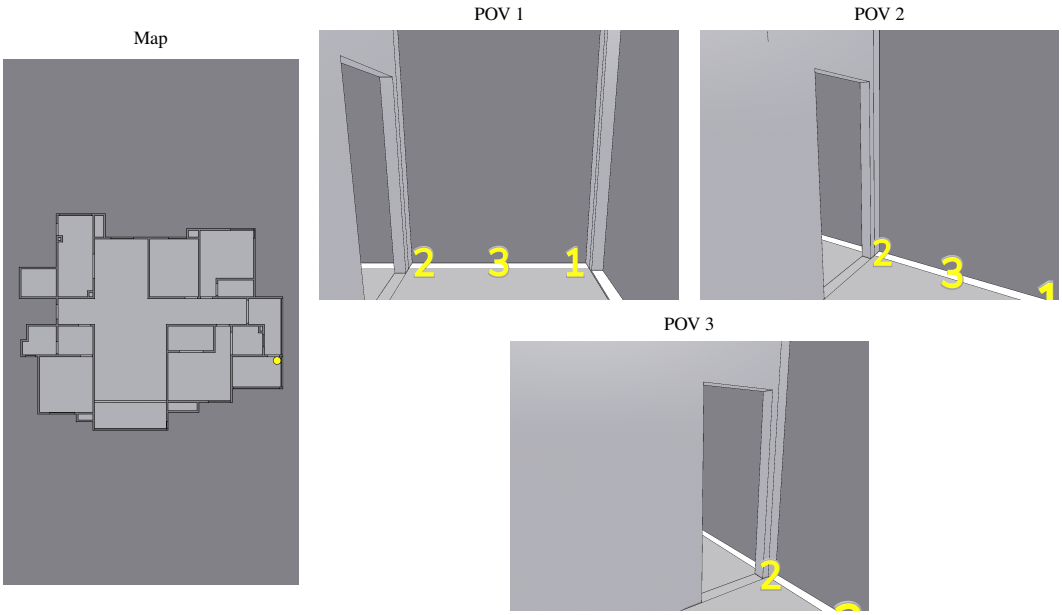

Figure 16: FLOORPLAN easy example. The yellow map marker identifies a corner location. The correct answer is **2** among 3 distinct numbered locations. The bin contains 1,000 questions.

Medium: 6 candidates, side target

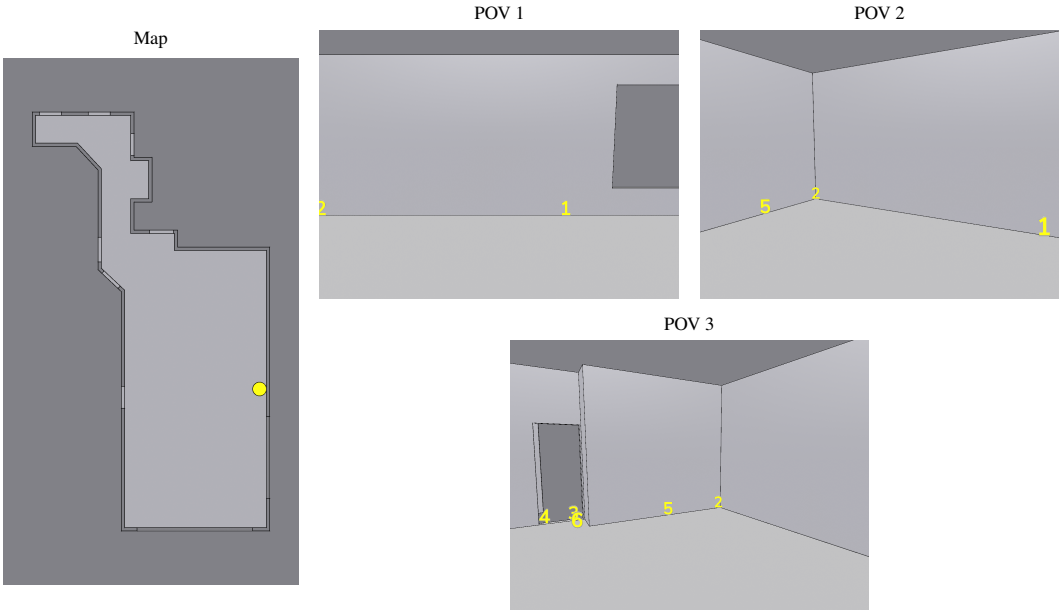

Figure 17: FLOORPLAN medium example. The yellow map marker identifies a side-midpoint location. The correct answer is **1** among 6 distinct numbered locations. The bin contains 1,000 questions.

FLOORPLAN EXAMPLES: HARD AND VERY HARD

Hard: 12 candidates, corner target

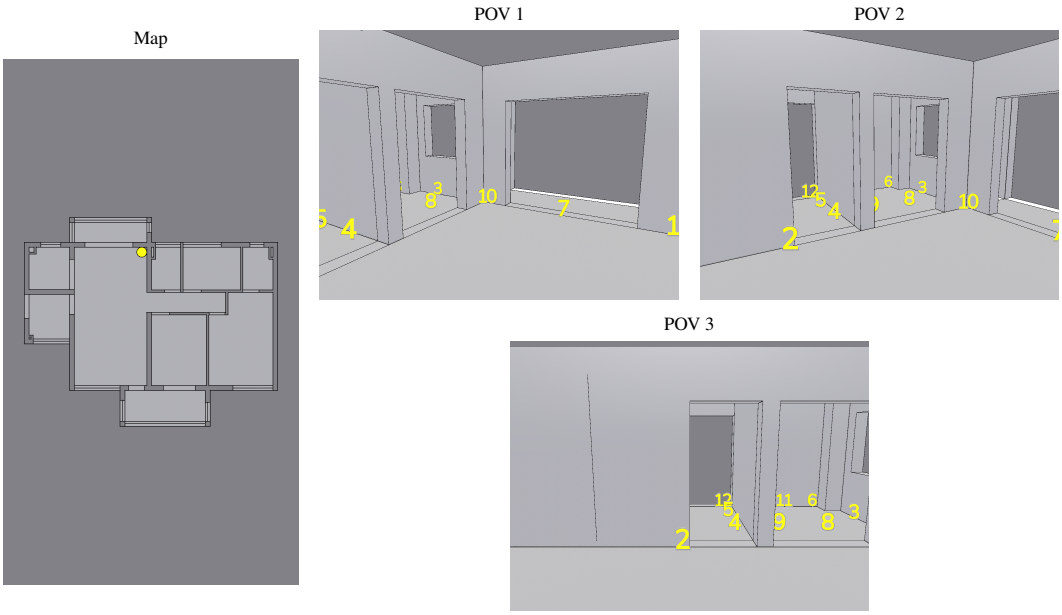

Figure 18: FLOORPLAN hard example. The yellow map marker identifies a corner location. The correct answer is **1** among 12 distinct numbered locations. The bin contains 1,000 questions.

Very hard: 13 candidates, side target

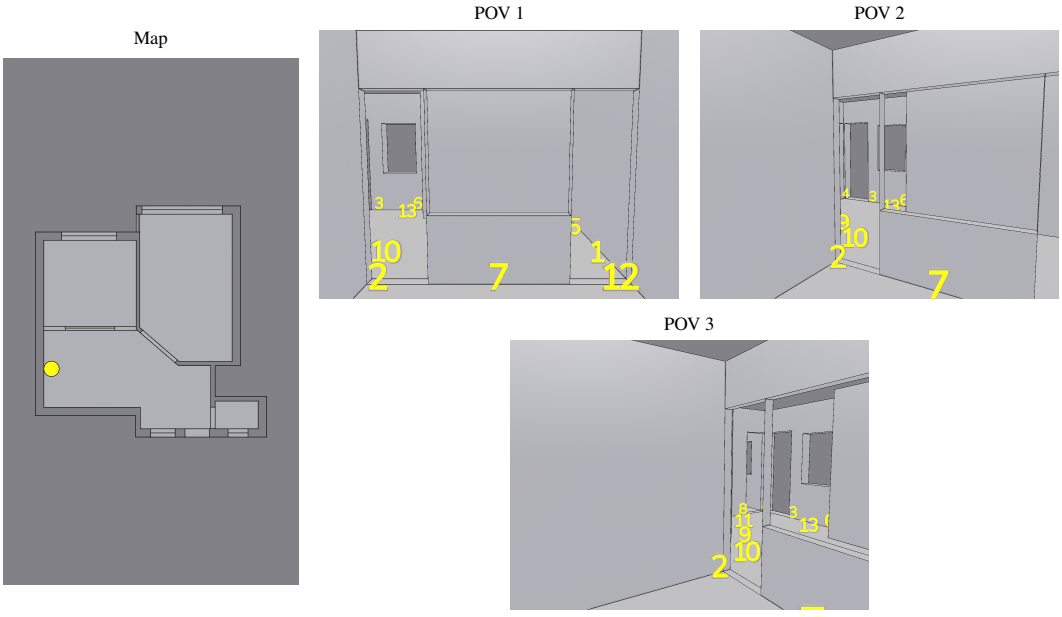

Figure 19: FLOORPLAN very hard example. The yellow map marker identifies a side-midpoint location. The correct answer is **5** among 13 distinct numbered locations. The bin contains 603 questions.

B HUMAN EVALUATION AND DATA COLLECTION

B.1 PARTICIPANT PROTOCOL

Recruitment and compensation. Both studies were run on Prolific. Eligibility required fluent English and an age between 21 and 40, and participants had to pass the screening questions described below. Participants were paid a base rate of $12/hour, with a $0.80 bonus for high performance. Participants gave informed consent in the study application. The study was reviewed by an institutional review board and determined to be exempt.

Session structure. Each participant completed one fixed question subset in a web application (Section B.2 gives the instructions). A session consisted of the instruction screen, a demo item (TRIANGLE only), two screening questions, and the scored questions. The application records the selected answer, correctness, response time and a per-question difficulty rating.

B.2 PROMPTS

Below we provide the benchmark-specific task instructions shown to participants.

TRIANGLE INSTRUCTIONS

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 in 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.

FLOORPLAN INSTRUCTIONS

You will see four images for each question.

The first image is a top-down floor plan of an indoor scene. A yellow circle marks the target location.

How to read the floor plan:

- The light gray areas are floor space / rooms.
- The darker outlined bands are walls or room boundaries.
- Gaps or breaks in walls usually indicate openings, such as doors or passages between rooms.
- Thin rectangular features along walls may indicate windows, doors, or wall openings.

The next three images show what the same scene looks like from first-person viewpoints, as if you were standing inside the rooms and looking around. Yellow numbers mark possible locations in the scene.

Your goal is to choose the number that matches the yellow circle on the floor plan.

Important details:

- The same number can appear in more than one first-person image. If it does, it means the same physical location seen from different angles.
- The numbers are reassigned for each question, so do not rely on numbers from earlier questions.
- Use the room layout, walls, corners, openings, and relative positions to match the floor plan to the first-person views.
- Exactly one numbered location is correct.

Select the number that best corresponds to the yellow circle.

C Model Evaluation Details

Access and timing. The commercial models were evaluated through provider-hosted API. All open models were evaluated from their public Hugging Face checkpoints.

Model-facing prompt and image order. Every model receives the same fixed prompt with four images attached in a fixed order: the map first, then the three egocentric views.

> Image 1 shows an overhead view of a scene, with the target location shown as a yellow circle. Images 2, 3, and 4 show overlapping first-person views of the scene with numbers that correspond to different locations. The same number appearing in different images refers to the same physical location in the scene. Which numbered location in images 2, 3, and 4 best matches the target location in image 1? Choose only one number.
>
> Return exactly two lines and nothing else (no markdown, no extra text):
> Explanation: ⟨your reasoning on a single line⟩
> Location: ⟨one number⟩

Image delivery. Images are submitted at their render resolutions (Triangle: 1920×1080 views, FloorPlan: 1440×1080 views; maps 1080×1920 in both). Local models apply their own official preprocessors.

Decoding settings. Local-model sampling settings follow recommended settings in the original work. For the Gemini models, we use the dynamic thinking mode for each model. Full benchmark results use three runs per model except GPT-6 Astra, which uses one run.

Answer parsing. A fixed parser extracts the answer from the raw response. It is applied identically to every model and seed:

- Markdown code-fence markers are stripped. `<think>...</think>` blocks are removed before parsing.
- If the response is a single JSON object, its `explanation`/`location` fields are read directly.
- Otherwise the parser takes the *last* occurrence of a `Location:` label followed by an integer; text before it is treated as the explanation.
- A response whose parsed location is missing, non-numeric, or not among the question's candidate labels is scored as incorrect.

Inference settings. Open-model runs used NVIDIA A6000 and B200 nodes. API runs used the providers' batch endpoints.

D Additional Results and Diagnostics

Table 7: Composition of the REMAP benchmark (7,491 questions total). Triangle crosses four angle configurations with a congruency flag (*Congr.* means the map and the views use the same representation format). FloorPlan is balanced over candidate-count bins (*Range* = number of visible candidate labels) and over the two target types, *Corner* (at a room angle) and *Side* (at a wall midpoint).

Triangle (3,888 questions)

Split	Angles (°)	Congr.	Incongr.
Low	(50,60,70)	324	648
Medium	(40,60,80)	324	648
High	(35,60,85)	324	648
Highest	(30,60,90)	324	648
Total		1,296	2,592

FloorPlan (3,603 questions)

Cand. bin	Range	Corner	Side
Easy	2-4	500	500
Medium	5-8	500	500
Hard	9-12	500	500
Very hard	13+	302	301
Total	2-23	1,802	1,801

E INSTRUCTION PROMPTS AND DIAGNOSTIC PROTOCOL

Figure 4 uses the 576 TRIANGLE questions in the human study. The sample has 64 questions per map/view cell, giving 192 Congruent and 384 Incongruent questions (128 of each in CORNERS/SIDES cells). Four models answer every question under Original and under the three added prompts below: GPT-6 Astra, Gemini 3.8 Flash, Gemini 3 Flash and Qwen3.8-27B. All conditions use the same four images in their original order.

Humans did not receive these prompts.

Shared Original prompt. Image 1 shows an overhead view of a scene, with the target location shown as a yellow circle. Images 2, 3, and 4 show overlapping first-person views of the scene with numbers that correspond to different locations. The same number appearing in different images refers to the same physical location in the scene. Which numbered location in images 2, 3, and 4 best matches the target location in image 1? Choose only one number.

Shared closing:

```
Return exactly two lines and nothing else (no
markdown, no extra text):
Explanation:  <your reasoning on a single line>
Location:  <one number>
```

Original contains only the shared opening and closing. Each intervention inserts one paragraph between them, separated by blank lines. Each addition contains exactly 45 whitespace-delimited words, not necessarily the same number of tokens.

Neutral (generic instruction). Please examine all four images carefully and take sufficient time to compare the available visual details before deciding on your answer, checking your initial impression against the evidence in each image and reviewing your reasoning once more before making your final choice of a number.

Attention to drawing styles. Please inspect the drawing styles in the overhead map and first-person views, paying particular attention to whether visible wall segments occur around corners or along sides. Check the depiction used in each image carefully before deciding which numbered location best matches the yellow target circle.

Explicit convention. The overhead map and first-person views may depict different geometric abstractions of the same underlying triangle: its full perimeter, segments around its corners, or segments centered on its sides. These representations preserve the underlying triangle and target locations even when different wall segments are shown.