

EgoShift: Bounding the Action-Only Shortcut in Egocentric Spatial Updating for Vision–Language Models

Faisal Ahmed John Quarles Kevin Desai
Department of Computer Science, University of Texas at San Antonio
`faisal.ahmed@utsa.edu`

Abstract

Robots and augmented-reality systems must track objects as their cameras move. A model may name an object correctly yet place it in the wrong direction after the camera turns. Recent benchmarks test this ability. They report one pooled accuracy per model, and guard against guessing with balanced answers, human references and text-only controls. By that measure the strongest models look close to solving the task. We show that a pooled score hides a failure those guards do not catch. These benchmarks ask where an object will be after a stated camera turn, and that turn alone rules out some answers. A model can therefore score well from the turn and never use the image. We call this an action-only shortcut. We present EgoShift, a benchmark of 12,786 questions in simulated indoor scenes. A model must match an object across two views, say where it lies after a stated turn, and judge whether it is visible, out of view, or hidden. Our main contribution is a test for the shortcut. The camera geometry determines which answers each turn can produce. We compute the highest score the shortcut can reach, then build every turn to sit exactly there. The test catches what the usual guards miss. An earlier version of our data looked balanced, yet a system that saw only the stated turn scored 72.3%. Across 36 models, accuracy does not fall as the turn grows larger. At 90° and 180° no frontier configuration scores above 50%, and most give the same answer they give with no image. Trained annotators score far higher. A benchmark whose questions state an action should report accuracy per action. A pooled number reports progress that has not happened. The benchmark, model responses and audit code are public at <https://github.com/UTSA-VIRLab/EgoShift>.

1. Introduction

A teapot sits on the table in front of you. Turn 90° to the right and it is on your left. The teapot has not moved; you

have. Robots and augmented-reality assistants do this constantly. They must also tell whether an object has left the camera’s view or is hidden behind something else. Figure 1 shows an example from EgoShift, where a 90° turn moves a cart from the right of the view to straight ahead.

Several recent benchmarks test this ability [17, 20, 38]. They report one pooled accuracy per model, and they use the usual controls against guessing: balanced answer choices, a text-only arm, and a human reference score. By these measures the strongest vision–language models now look close to solving the task. We find that the pooled score hides a failure those guards do not detect.

The cause is a shortcut in questions that name an action. When a question states how far the camera turned, that turn alone makes some answers likelier than others. A model can then answer well from the action and never look at the image. We call this an *action-only shortcut*. This shortcut is a form of the language priors reported in visual question answering [9, 11]. It is also the kind of leak that benchmark designers are told to remove before release [3]. Answer balancing does not remove it. The imbalance is inside each turn, not across the whole question set. We address it with a number we compute in advance and then build the benchmark to match.

EgoShift is a benchmark of 12,786 held-out questions in simulated indoor scenes built with AI2-THOR and Proc-THOR [7, 18]. It has three tasks (Fig. 1), with 5,838, 1,098 and 5,850 questions respectively. Task A shows two views and asks whether a marked object is the same one. It is a control. Task B shows one view, states a camera turn, and asks where the object will be afterwards without showing the result. This is the primary task. Task C asks whether an object seen earlier is still visible, has left the view, or is hidden behind something else.

For each turn angle and direction in Task B, we compute the highest score a predictor can reach from the action alone. Five angles and two directions give ten such groups, and we build each one to have exactly that score (Sec. 3.3). The value follows from the geometry, and it is fixed before any model runs, so it cannot be adjusted to suit a result.

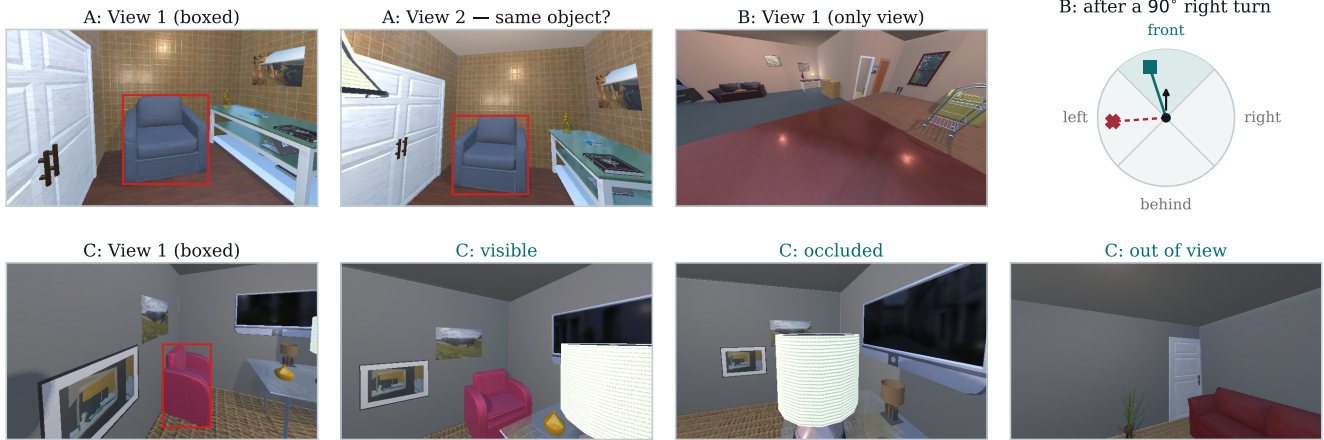


Figure 1. Examples from EgoShift, all from the development split. The red rectangle is the *box*, which marks the object in question. On the Task-B dial, teal is the correct answer and red is the answer a model gets by assuming the cart started straight ahead. In the Task-C occluded view a desk lamp 0.80 m from the camera covers the chair. Walls, doors and windows are never allowed to be the covering object.

This check is not redundant with the usual ones. An earlier version of our data was balanced by every conventional measure, yet an action-only predictor still scored 72.3% on it (Sec. 3.3).

Perspective change, spatial updating, occlusion, human references and text-only baselines all appear in recent benchmarks, SCOPE most directly [17]. None of them derives the exact score the stated action alone can reach, turn by turn, before any model is evaluated. EgoShift adds that derivation, and it lets us report accuracy one action at a time. The recipe is not specific to EgoShift. Any benchmark whose questions state an action can compute the score reachable from that action alone. It can then build against that score before balancing. In summary, we make three contributions.

1. **A construction test for the action-only shortcut.** We define the shortcut precisely, derive the best score it can reach for each turn, and require the built data to match that value. We release the derivation, an independent checker, and a record of which scenes were ever used during development. The test found a 72.3% shortcut in data that conventional answer balancing had passed.
2. **EgoShift.** A three-task benchmark of 12,786 questions covering cross-view identity, direction after a stated turn, and visibility. Development and test questions come from separate rooms, and all three tasks carry a trained-human reference.
3. **A per-action analysis of 36 models.** We show that pooled accuracy hides a failure confined to 90° and 180°, where no frontier configuration scores above 50%. Nearly every frontier error matches one rule: a model that assumed the object lay straight ahead, then applied the stated turn, would give the same answer. At those two turns the models return their no-image answer. People find those turns hard too, but never fall to that score. Raising the reasoning budget does not help.

2. Related Work

Keeping benchmarks from being solved the wrong way. CLEVR [15] rendered every scene programmatically, so an answer can be checked against the known scene state rather than against a label. GQA [12] and VQA v2 [9] then addressed a different problem: a model can answer from the question alone. They used scene graphs and paired images to even out the answer priors. Work on question-only bias [1], annotation artifacts [11] and shortcut learning [8] showed how general the problem is. Adversarial filtering [19] responded by training a model to find the shortcut, then removing the questions it catches. TsT/IBP [3] makes the same argument for benchmark designers today: train blind predictors on the fields a benchmark publishes, then remove the questions they exploit. How far it goes depends on how well a learned predictor does.

EgoShift instead bounds one shortcut that can be written down in advance: answering from the stated camera turn with no image at all. We declare the camera geometry, so the best score that shortcut can reach follows from the geometry and not from any model. We hold each turn to that score. The two approaches answer different questions and we use ours alongside learned attacks, not in place of them.

Exact labels are not enough on their own. Work on unanswerable questions [10] and on re-annotating test sets [2, 26, 28] shows that items which are ambiguous to people can move both scores and rankings. We therefore check labels against the simulator and check answerability against raters, and report what the raters scored.

Spatial benchmarks, and what they already cover. Early diagnostics found that models fail on static spatial relations [16, 22]. Later suites widened the target to metric 3D, perspective taking and multi-task spatial reasoning [14, 24, 31, 37], and SpatialScore alone spans 30 tasks [33]. EgoShift is far narrower and claims no broad coverage.

OSI-Bench [34] and Unreal3DSpace [25] sit at opposite ends of simulated indoor evaluation. OSI-Bench moves out to the open world, and Unreal3DSpace controls the scene closely. We stay inside simulation because it gives us exact labels for the camera’s own motion.

Turning the camera, and matching across views. Several benchmarks already ask what a scene looks like from somewhere else. ViewSpatial-Bench [20], SAT [27], Spin-Bench [41] and CVT-Bench [30] cover perspective taking, rotation, and viewpoints whose image is withheld. SCOPE [17] covers updating and occlusion across eight viewpoints, with a text-only baseline, a human reference and correlations with embodied evaluations. PCSR-Bench is the closest benchmark to ours by task, with 84,373 questions on egocentric rotation, re-anchoring and visibility under a limited field of view [4]. Ego-ST Bench extends egocentric evaluation to video [35]. We therefore do not claim viewpoint updating or egocentric rotation as new.

The same holds for the other two tasks. Multi-image suites test what a model builds from limited views [38, 39]. Cross-view work covers correspondence, visibility and occlusion in real imagery [23, 29, 32, 43]. TriViewBench varies object count and occlusion in controlled synthetic scenes [5], and MindEdit-Bench asks what happens when an object moves [40]. Task A is accordingly a control and not a contribution. Task C differs from MindEdit-Bench in what moves: our camera turns while the room stays still.

Models built for spatial reasoning. A separate line of work improves the models rather than the benchmarks. It includes multi-frame training [36], cross-view supervision that transfers to manipulation [13], reinforcement learning aimed at transformations [21], latent 3D imagination [6] and camera-guided fusion [42]. SITE and SCOPE relate benchmark scores to embodied evaluations [17, 31]. For this reason our panel includes spatially tuned checkpoints. We report no downstream result of our own.

Where that leaves EgoShift. EgoShift introduces no new spatial task. SCOPE [17] and PCSR-Bench [4] are the closest by task. SCOPE reports updating and occlusion across eight viewpoints, and PCSR-Bench reports egocentric rotation and limited-field-of-view visibility. EgoShift adds a construction target instead. We compute the highest score the action-only shortcut can reach and hold every turn to it. We then test the items separately for other shortcuts, and for whether people can answer them. The supplement compares EgoShift with the closest existing benchmarks, one by one.

3. Building EgoShift

EgoShift holds 12,786 held-out questions across 4,698 rooms, with a further 15,761 released for development. Table 1 of the supplement gives the composition of each task.

3.1. Scenes and camera

Scenes are furnished indoor rooms generated with AI2-THOR and ProcTHOR [7, 18] and rendered at 640×480 . The rooms are simulated, so they contain no people and no personal information. We restrict the camera, and we restrict it most for Task B. Every Task-B image is taken at standing height, with a constant 30° downward tilt and a 120° vertical field of view. The stated turn is a rotation about the vertical axis, made from that same spot. Only the heading changes. That is why the construction test in Sec. 3.3 is possible: the set of answers a turn can produce is then finite, and we can compute it exactly. Tasks A and C show two views, and for those the camera also moves across the floor. Neither task states a turn, so nothing in the derivation depends on their poses.

The rooms are simulated, so we do not read the answer off the image. The simulator reports which pixels belong to which object, where each object sits in three dimensions, and where the camera is pointing. We re-render every saved pose and take the answer from that record rather than from the stored scene description. The re-render caught an earlier build that recorded a 90° field of view for 120° frames.

3.2. The three tasks

The three tasks answer different questions and are never combined into a single score. The held-out sets draw on 3,119, 736 and 1,334 rooms and cover 28, 45 and 49 object categories. Task B comes last, because its question names an action and needs the derivation in Sec. 3.3.

All three ask a multiple-choice question with a fixed answer set, so scoring needs no judgement about free text. Each task uses a single question template. Only the object named changes, and for Task B the size and direction of the turn. We shuffle the answer options in every question, so a model that always picks the first option scores at the chance rate. In the held-out sets each answer appears first about as often as expected. The largest deviation is 2.2 percentage points. Task A also shuffles which image is View 1.

Task A, the control: is it the same object? Two images of one room each show a rectangle drawn around an object of the same kind, for example two chairs. We call this rectangle the *box*. It tells the model which object the question is about. The model answers whether the two boxes contain the same physical object or two different ones. A target must show at least 2,000 pixels and a tenth of its silhouette in both views, so neither box contains a fragment too small to identify.

Task A checks that a model can find a boxed object and match it across two views under this renderer. A low Task-B score therefore cannot be blamed on the imagery.

An earlier build had a cue that we removed. Targets fully visible in both views were nearly always *different*. A *different* question shows two fully visible objects. A *same* ques-

tion shows one object twice, and the second view usually cuts part of it off. We select these questions to remove that cue. We match *same* against *different* within object type, and within bins of smallest coverage, largest coverage and distance of the box from the image centre. Every object type then has equal numbers of both answers. The class prior and object type alone therefore each score 50%, and the two answers end with almost the same coverage. When we withhold the images, or give only one, the frontier configurations score at the chance rate (Sec. 4.2).

Two things remain unsettled. A finer selection matches on continuous statistics rather than bins, and gives a smaller question set that we have not adopted. We have also not run a classifier trained on the remaining visibility statistics. Both belong with the learned attacks in Sec. 5.

Task C: can you still see it? The first image boxes a target. The second shows the same room after the camera has turned. The model answers *visible*, *out of view*, or *occluded*. We calculate the correct answer from the simulator. A target counts as visible when at least 300 of its pixels are drawn, about a tenth of a percent of the frame. That threshold comes from the raters. An earlier build used 30 pixels, and every visible-class error the panel made was on a question showing fewer than 300. When no pixels are drawn, the target is out of view if its whole three-dimensional extent lies outside the region the camera can see. It is occluded if it lies inside that region and a nearer object covers it, and that object may not be a wall, door or window.

Task C questions come in matched triples. For each target we build one question per answer. All three share the same first image and use different second images, so the held-out set has exactly 1,950 questions per answer. Every target appears once with each answer. A model that guesses the most common class therefore scores 1/3, and so does a model that recognises the target and recalls its usual answer. That is a property we build in, not a claim that Task C has no other shortcuts.

Task B, the primary task: where is it after you turn? The model sees one image and a question naming the target by kind, for example *the sofa*. We draw no box. On all 1,098 held-out questions only one object of that kind is visible, so the name is enough. The question states an in-place turn of 30°, 60°, 90°, 120° or 180°, left or right. It then asks where the object will be: in front, behind, to the left, or to the right. We never show the view after the turn. A model must take the object’s direction in the image and update it by the stated turn.

Three conditions make the answer unambiguous, and all three hold on every held-out question. The target shows at least 2,500 pixels, median 3,807, and at least 80% of its silhouette. All eight corners of its three-dimensional box land in the same answer after the turn, not merely its centre. The nearest corner is at least 15° from the boundary between

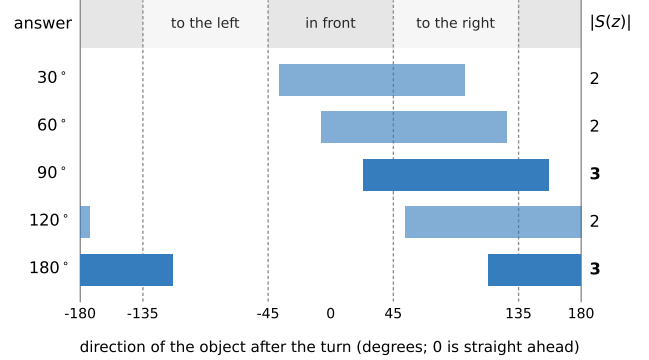


Figure 2. Why some turns allow three answers and others allow two. Dashed lines are the boundaries between the four answers, at $\pm 45^\circ$ and $\pm 135^\circ$. Each bar is the band of directions the object can occupy after a left turn of that size. The bands start from the 66.5868° band that construction allows. Right turns mirror it.

two answers, median 26° , so no question is a near-tie. The questions are also spread widely. No two share a target object or a source image, no room contributes more than two, and no category accounts for more than 55 of the 1,098.

3.3. Bounding the action-only shortcut

The turn stated in a Task-B question rules out some answers on its own. A system that never sees the image can therefore score better than chance by answering the most likely direction for each turn. We call this the *action-only shortcut*. Tasks A and C state no action, so they have no equivalent shortcut, and the balancing above protects them instead. We make the shortcut explicit, compute the highest score it can reach, and hold the built data to that score.

The shortcut, stated precisely. Write $z = (r, d)$ for the turn stated in the question, where r is its size and d its direction. Five sizes and two directions give ten such pairs, and we call each one a *cell*. Consider a predictor given z and nothing else, not the image, the scene, the object, the camera pose or the wording. For each z , its best strategy is to answer whichever direction is most common among that turn’s questions. We call its accuracy the *action-only ceiling*, and we want it as low as the geometry allows.

Let $S(z)$ be the set of answers cell z can actually produce, given the range of starting directions construction allows. Write $p(z)$ for the fraction of questions in cell z . If answers were spread evenly over $S(z)$, the ceiling would be $C_{\text{even}} = \sum_z p(z) / |S(z)|$. A finite set of questions cannot spread them perfectly. Let n_z be the number of questions in cell z , and let $N = \sum_z n_z$ be the total. In each cell, some answer must then occur at least $\lceil n_z / |S(z)| \rceil$ times. The lowest score the shortcut can be held to is therefore

$$C = \frac{1}{N} \sum_z \left\lceil \frac{n_z}{|S(z)|} \right\rceil. \quad (1)$$

Turn z	questions n_z	answers $ S(z) $	cell ceiling	commonest answer	met
30° left	135	2	68	68	✓
30° right	135	2	68	68	✓
60° left	135	2	68	68	✓
60° right	135	2	68	68	✓
90° left	117	3	39	39	✓
90° right	117	3	39	39	✓
120° left	134	2	67	67	✓
120° right	134	2	67	67	✓
180° left	28	3	10	10	✓
180° right	28	3	10	10	✓
Total	1,098	—	504	504	10/10

Table 1. The Task-B held-out set, one row per turn. The ceiling $\lceil n_z/|S(z)| \rceil$ is the fewest questions the most common answer can cover, and “commonest” is how many it does cover. They match in every row, holding the shortcut to $504/1,098 = 45.90\%$. Few questions have 180° geometry, so those rows are small (Sec. 3.3).

A build meets this ceiling only under two conditions. Every answer that occurs must lie in $S(z)$, and the most common answer in each cell must occur exactly $\lceil n_z/|S(z)| \rceil$ times. Answer balancing across the whole set does not achieve either, because the imbalance that matters is inside individual cells.

Which answers a turn can produce. $S(z)$ depends on how far off-centre the target may start. Construction requires the target’s centre to lie within 66.5868° of straight ahead. We did not choose that figure. It is the horizontal half-angle of a 640×480 frame at a 120° vertical field of view, $\arctan((640/480) \tan 60^\circ)$. It was written into the build before the poses were replayed. Under it, the numbers of reachable answers at 30° , 60° , 90° , 120° and 180° are 2, 2, 3, 2 and 3, for either direction. Fig. 2 shows why. A turn shifts the whole band of starting directions around the circle, and the band covers three answers only at 90° and 180° . A larger turn does not widen the band, so a larger turn is not a harder question.

Two independent implementations compute $S(z)$: an interval derivation, and an enumerator that sweeps 200,000 starting directions. They share configuration values but no code. We run both over ten camera configurations, 96 cells in all, and they agree on every one. Ours is one of the ten. Six of the rest vary the width of the starting band, from 45° to 75.1° , or rotate the answers by 15° . Three vary the turn set. The width matters most, because it is the one quantity we picked and the derivation has to change with it. It does change: at 45° only one answer is possible in four of the ten cells, and the ceiling for an equal mixture of cells rises to 70%. At 75.1° it drops to $1/3$.

The held-out set meets the ceiling in all ten cells (Tab. 1),

holding the shortcut to $504/1,098 = 45.90\%$. We use that figure throughout because it counts every question once. The ten per-cell rates average 43.96%, and C_{even} is 45.60%.

Two earlier builds passed conventional balance checks and failed this test. The first was balanced on every conventional measure: the most common answer overall scored 25.0%, and so did the most common answer per turn direction. Yet a predictor using turn size together with direction scored $833/1,152 = 72.3\%$, because some cells held only part of the answer set their geometry allowed. The second was balanced to 43.33% by the even-spread formula. When we replayed its poses, 304 of its 988 source rows violated the centring condition that formula assumes. Balanced margins are not enough, and a calculation from geometry is only as reliable as the assumptions we check it against. We evaluate neither build anywhere in this paper.

The same geometry explains why the cells differ in size. The two 180° cells hold 28 questions each, against 117 to 135 elsewhere. Those answers need the target to start between 45° and the edge of the permitted band, and the 15° boundary margin leaves only about 6.5° of that usable. We mined the available questions first, and the supply set the cell sizes. The build recorded the smaller cells and recomputed Eq. (1) before any model ran, so the ceiling is specific to this build and could not be adjusted afterwards.

3.4. Splits

Questions used for development and questions held out for testing never come from the same room. We match rooms by content rather than by filename, so a room saved under two names counts once. The 4,701 identifiers reduce to 4,698 distinct rooms, and no room appears in both roles.

We also track which rooms we ourselves have looked at. A record marks 3,715 rooms whose outcomes we saw during evaluation, annotation, pilot analysis or inspection by the authors. We draw held-out questions only from the remainder, and the audit reports zero seen rooms in every held-out split. This accounts for our own history with the data, not for what a model may have seen in pretraining.

Availability and licence. The public artifact at <https://github.com/UTSA-VIRLab/EgoShift> is reachable without contacting us. It holds the questions for every split, the development answers, the construction, scoring and audit tools, and every raw model response behind our tables. A reader can therefore recompute our numbers instead of trusting them. We withhold only the held-out answer key, which keeps the benchmark usable as a test. The artifact ships Croissant metadata with the Responsible AI fields. Our annotations, questions and audit outputs use the CC-BY-4.0 licence. We redistribute the rendered frames under the upstream Apache-2.0 terms. We audited all 9,838 staged houses, and all 1,465 distinct object assets in them come from the AI2-THOR library. No artifact contains a

Model	A	B	C
Gemini 3 Pro (thinking)	91.4	83.6	88.0
Gemini 3 Flash	96.6	83.5	89.3
Gemini 3 Pro	91.8	81.9 [†]	88.2 [†]
Claude Fable 5	94.8	81.4	88.4
GPT-5.6 Sol (thinking)	93.2	81.1	88.0
Claude Opus 5 (extended)	95.1	80.2	84.3
Claude Opus 5	91.6	80.1	83.2
Claude Sonnet 5	92.6	77.1	75.8
GPT-5.6 Sol	93.4	76.3	87.2
Gemini 3 Flash Lite	96.2	53.0	67.3
Claude Haiku 4.5	71.2	45.4	38.8
GPT-5.1	90.8	42.0 [†]	73.1
InternVL3-38B	96.1	65.1	65.2
Qwen3-VL-32B	95.6	47.7	59.3
SenseNova-SI-1.3-Qwen3-VL-8B	93.6	33.3	61.7
Qwen2.5-VL-7B	60.8	18.9	41.4
Nominal random	50.0	25.0	33.3
Action-only ceiling	–	45.9	–

Table 2. Main-arm accuracy (%). Twelve frontier configurations above the rule, four open-weight systems below. All 36 are in the supplement. Parse failures count as incorrect, as do the [†] cells, where a retry returned a second valid answer on six GPT-5.1 Task-B and on twenty-three and ten Gemini 3 Pro Task-B and Task-C questions.

human response file, a participant name or a pseudonym map. The supplement’s datasheet gives the full contents and the release plan.

4. Model and Human Results

We evaluated 36 systems (Tab. 2). Twelve are configurations of current proprietary models. Three of those twelve are the same model rerun at a raised reasoning setting. The rest are open-weight models and checkpoints tuned for spatial reasoning. Where we compare open-weight models by size, *larger* means 32 to 78 billion parameters and *smaller* means 2 to 14 billion. We score Task A on a fixed 1,192-question subset, Task B on its complete 1,098-question held-out set, and Task C on a 1,200-question subset drawn as whole matched triples. Per-class accuracy for all 36 systems is in the supplement.

4.1. Accuracy does not fall as the turn grows larger

Over the twelve frontier configurations, Task-B accuracy averages 86.8% at 30°, 82.1% at 60° and 84.3% at 120°, but only 38.3% at 90° and 36.8% at 180° (Fig. 3). The 95% scene-clustered bootstrap intervals over 2,000 draws are [85.5, 88.1], [80.0, 84.3], [82.3, 86.3], [32.8, 44.1] and [25.9, 48.8]. The 120° and 90° intervals do not overlap, and no configuration exceeds 50.0% at either hard turn. The 180° cell holds 56 questions, so we also report the two hard turns together: 38.0% over 290 questions, interval [32.9, 43.1]. A 120° turn is answered correctly more than twice as often as a 90° one, so a single pooled score hides this difference entirely.

People find the same two turns harder, so the shape is

not only about models. On the 120 questions the panel and the models share, rater accuracy falls from 93.1% at 30° to 74.1% at 90° and 77.8% at 180°. The action-only ceiling separates them. Raters stay 41 to 42 percentage points above that ceiling at both turns. The model average comes within 5.0 percentage points at 90° and within 1.1 at 180°.

Models answer on the wrong axis rather than near the right answer. At 90° Claude Opus 5 answers 100.0% of the *to the left* and *to the right* questions correctly, against 1.3% of *in front* and 2.6% of *behind*. At 180° the pattern reverses, and Gemini 3 Flash and Claude Fable 5 behave the same way. The turn stated in the question decides which axis is wrong.

One rule reproduces that pattern. We call it the *straight-ahead rule*. The model ignores where the target actually is, assumes it lies straight ahead, and then applies the stated turn correctly. It predicts $\text{zone}(\delta)$ for a signed turn δ . It matches 98.7% of Claude Opus 5’s answers at 90° and 100.0% at 180°. It agrees with the correct answer only 33% of the time there, and about 50% elsewhere. We rejected a simpler account, that these models never update at all. A model that answered in the un-rotated starting frame would score 25.0%, and that rule matches what the models said on only 27–33% of 90° questions. The pattern therefore fits a rotation applied to an assumed direction rather than to the one in the picture. Section 5 states what this evidence does and does not settle.

Fig. 4 shows the substitution on a mirrored pair. There are 127 held-out questions at 90° and 180° where the five strongest configurations all give the assumed-ahead answer.

At 90° and 180° the picture does not change the answer.

We can also test this without the rule. For each model we compare its answer with its own no-image answer. Across the twelve frontier configurations the two agree on 59.4% of 30° questions, 63.0% at 60° and 65.6% at 120°, against 94.8% at 90° and 89.9% at 180°. The intervals are [48.9, 70.1], [48.1, 78.1], [50.8, 79.5], [91.3, 97.9] and [86.5, 93.2], from 24 questions per turn per model. This compares a model with itself on identical questions, so neither model ranking nor question difficulty explains the gap.

Two kinds of error cover almost every frontier mistake.

An error is *assumed-ahead* when the answer is the one the rule predicts. It is *single-axis* when the model answers on one axis for more than 80% of questions and errs there. Together they cover 96.3% of the 3,670 frontier errors. Assumed-ahead alone accounts for 82.0%, and for 86.9 to 100.0% of the errors made by the eight strongest configurations. These systems are not making many different spatial mistakes. They are making one mistake repeatedly.

The proportions reverse for the smaller open models, where single-axis errors dominate at 44.0% and assumed-ahead explains 27.4%. Held-out answers are 47.3% left-or-

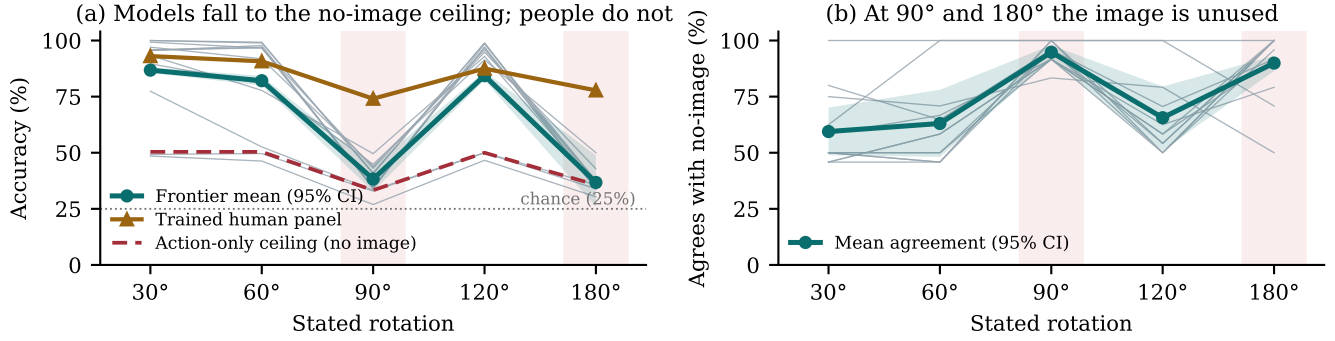


Figure 3. (a) Task-B accuracy against stated rotation. Thin lines are the twelve frontier configurations. (b) The percentage of answers identical to the same model’s answer with no picture.

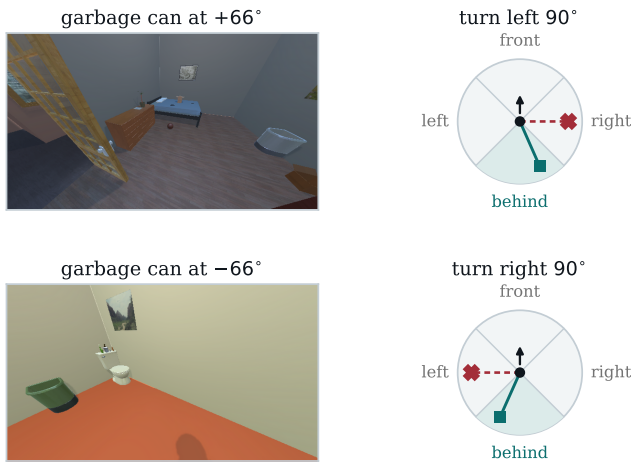


Figure 4. Two held-out questions where the five strongest frontier configurations all gave the same wrong answer. The correct answer is *behind* in both (teal); all five answered the side zone (red). The mirroring rules out a sign error in our geometry.

right and 52.7% front-or-behind, so a model answering almost entirely on one axis is capped well below chance. That is why eleven of the 36 systems fall below the 25% random rate without guessing randomly. It is also why GPT-5.1, which answers left or right on 93% of questions, ranks last in the frontier group. Per-model tables are in the supplement.

More reasoning does not help at the two hard turns.

Three configurations ran at a raised reasoning setting against their own default. All three already reason by default, so the contrast is one of degree. Over the whole task, GPT-5.6 Sol gains 4.8 percentage points, Claude Opus 5 changes by 0.1, and Gemini 3 Pro loses 0.3. At 90° and 180° all three score lower: -0.9 and -7.1, -0.4 and 0.0, -2.1 and -1.8. The extra computation goes to turns other than the two where the failure appears. That is what the straight-ahead rule predicts. Poor reasoning about a direction the model had already located would not predict it.

4.2. Controls and the human reference

What the models score with the images withheld. Every control arm runs on the same 120 questions the annotators saw. With both images withheld, the twelve frontier configurations average 46.1% on Task A against a 50.0% random rate, and 29.7% on Task C against 33.3%. No system exceeds 53.3% or 35.8%, against 71.2–96.6% and 38.8–89.3% for the same systems on the same items with the images shown. Two fall below the random rate, which we read as refusal rather than as guessing worse than chance. The twelve open models score 46.7–53.3% and 31.7–35.0% blinded, and one view leaves them at the chance rate. A model cannot answer Tasks A and C from the wording or from a single view.

Task B is different, because the turn a question states really does narrow the answers. All twelve frontier configurations score between 42.5 and 44.2% blinded, mean 43.4%, close to this subset’s action-only ceiling. A spread of 1.7 percentage points across twelve systems makes that a property of the questions, not of any model. A score above the 45.9% ceiling is therefore not evidence that a system used the image. The comparison in Sec. 4.1 settles that.

The trained-human reference. The institution’s review board determined before collection that the study does not constitute human subjects research requiring oversight. Nine paid lab annotators worked from a written instruction sheet with practice items and attention checks. They returned 25 complete sessions between 24 and 27 August 2026. One rater completed Tasks B and C on 29 August, after that date. We added both to the panel, which brings the total to 27 sessions. The supplement reports the effect. We pseudonymise responses on arrival. Mean rater accuracy is 98.1% on Task A, 84.6% on Task B and 89.6% on Task C. Panel majorities are 100.0, 96.7 and 91.7%, with Fleiss’ κ of 0.960, 0.681 and 0.820. Two-way cluster bootstrap intervals over raters and scenes stay well above the random rates. On Task B the lower endpoint of 77.3% clears the 43.4% action-only ceiling for this subset. We call this a

trained-human reference, never a human ceiling.

Models and raters share all 120 Task-B questions, where every frontier configuration scores between 41.7 and 75.8% against the panel’s 84.6%. For Tasks A and C the model subsets overlap the human questions on only 25 and 23 items, too few to compare. Raters find one Task-C class hardest. They reach 74.4% on occluded questions against 97.8% out of view and 96.7% visible. Three departures from the written protocol limit this reference, and the supplement lists them in full. We exceeded the five-rater target and kept every valid session, rather than selecting five after seeing scores. No rater followed the intended Task-B-first ordering, so that estimate may carry familiarity from the other tasks. Every Task-A session and six of eight Task-C sessions missed one shared attention-check item. We report this as a fault in the instrument rather than grounds for removing anyone.

How to report results on EgoShift. Report Task B one turn at a time and Task C one answer class at a time, and treat questions from the same room as dependent. Compare each score against the random rate, that turn’s action-only ceiling, and the trained-human reference. A score above the ceiling is necessary but not sufficient, because a system can reach it without using the image. Do not tune on the held-out questions. The supplement states the full protocol.

Where models still have room to improve. The pooled score puts Task B’s best configuration at 83.6%, against a trained-human mean of 84.6%. When we read the score one turn at a time, those two numbers stop matching. At 30°, 60° and 120° the best model beats the panel by 6.9, 8.6 and 11.4 percentage points, while at 90° and 180° it trails by 24.5 and 27.8. Progress on EgoShift should therefore appear as higher scores at those two turns. A system that raises its pooled score while leaving them at the action-only ceiling has not improved at what the benchmark measures. Task C has room to improve only in its occluded class. There the best model (88.0%) already exceeds the panel (74.4%), so we do not treat the human number as a bound either.

5. Limitations

Every scene is simulator-rendered at one resolution and one field of view, and the only action a question states is an in-place turn. The benchmark has no photorealism, no sensor noise, no changes in pitch and no multi-step navigation. Nothing here shows that the results transfer to natural images or to a robot. Object categories follow the AI2-THOR asset library, so neither the objects nor the room layouts represent real interiors. Task A is a control and carries no difficulty claim. The systems we evaluated are not a random sample, so comparisons between them are descriptive.

The ceiling in Sec. 3.3 bounds only one shortcut: answering from the stated turn with no image. A build that attains the ceiling has shown nothing about wording, meta-

data, or a cue that some future model finds and we did not anticipate. Learned attacks against the built questions are a separate check, with a protocol fixed in advance and stated in the supplement. We do not claim the questions are free of shortcuts in general.

The straight-ahead rule is an interpretation of the evidence, not an established mechanism. Three tests converge on it: the answer-axis pattern, the agreement with the no-image arm, and the failure of extra reasoning to help at 90° and 180°. We have not run the matched counterfactuals that would settle it. Those would vary the stated turn while holding the picture fixed. A model that behaved this way for another reason would produce the same evidence.

The human reference comes from a small trained lab panel of nine raters per task. It estimates trained task performance, not a bound on what people can do, and Sec. 4.2 lists three protocol deviations that limit it further. Raters also lose accuracy at 90° and 180°, so part of the difficulty at those two turns belongs to the questions rather than to the models. On the 120 questions raters and models share, the panel tied or disagreed with the simulator on four Task-B questions and on ten Task-C questions. All ten Task-C cases are in the occluded class. Rater accuracy without those questions rises from 84.6 to 86.3% on Task B and from 89.6 to 94.8% on Task C. The gap we report between raters and models is therefore conservative. We changed no question and removed none, and we collected no rater demographics, so we cannot say how representative it is.

6. Conclusion

We introduced EgoShift, a benchmark of 12,786 held-out questions in simulated indoor rooms. It asks whether a vision-language model still knows where objects are once the camera turns. A model must match an object across two views, place an object after a stated turn, and judge whether an object is visible, hidden, or out of frame. The benchmark’s contribution is how it is built. A question that states how far the camera turned can be answered from that turn alone, with no image. We declare the camera geometry, so we compute the highest score that shortcut can reach, turn by turn, before any model runs. We build every turn to sit exactly on that value.

Ordinary answer balancing does not achieve this. An earlier version of our questions passed every conventional check, while a system that saw only the stated turn scored 72.3%. Across 36 systems, accuracy does not fall as the turn grows larger. Accuracy drops sharply at 90° and 180°, where no frontier configuration scores above 50% and most return the answer they give with the images withheld. Trained annotators find those two turns harder too, yet stay far above that no-image score.

A pooled score hides all of this. When a benchmark question names an action, the score reachable from that ac-

tion alone can be computed in advance. That score should be reported beside the results. EgoShift, its derivation, an independent checker and every model response are public.

Acknowledgments

This work was supported by the National Science Foundation under award [AWARD NUMBER]. We thank the paid annotators of the San Antonio Virtual Environments (SAVE) lab for the trained-human reference.

References

- [1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018. 2
- [2] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiao-hua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. 2
- [3] Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should “train on the test set” to expose exploitable non-visual shortcuts, 2025. 1, 2
- [4] Yuangong Chen, Wai Keung Wong, Jiaxing Li, Ioannis Patras, and Xu Zheng. Beyond localization: A comprehensive diagnosis of perspective-conditioned spatial reasoning in MLLMs from omnidirectional images, 2026. 3
- [5] Yu-Yang Chen and Lan-Zhe Guo. TriViewBench: Controlled complexity scaling for multi-view structural reasoning in MLLMs, 2026. 3
- [6] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3D: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2624, 2026. 3
- [7] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3
- [8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2
- [9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Evaluating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. 1, 2
- [10] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, 2018. 2
- [11] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018, Volume 2 (Short Papers)*, pages 107–112, 2018. 1, 2
- [12] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019. 2
- [13] Suchae Jeong, Jaehwi Song, Haeone Lee, Hanna Kim, Jian Kim, Dongjun Lee, Dong Kyu Shin, Changyeon Kim, Dongyoon Hahm, Woogyel Jin, Juheon Choi, and Kimin Lee. Learning multi-view spatial reasoning from cross-view relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2570–2581, 2026. 3
- [14] 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. 2
- [15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. 2
- [16] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9161–9175, 2023. 2
- [17] 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)*, pages 32803–32827, 2026. 1, 2, 3
- [18] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 1, 3
- [19] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1078–1088, 2020. 2
- [20] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models, 2025. 1, 3

- [21] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. STAR-R1: Multi-view spatial TrAnsformation reasoning by reinforcing multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12041–12051, 2026. 3
- [22] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2
- [23] Ruixun Liu, Lingyu Zhang, Lanxuan Xue, Kaiyu Li, Bowen Fu, and Xiangyong Cao. CVSBench: A comprehensive benchmark for cross-view spatial reasoning and dreaming, 2026. 3
- [24] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSR-Bench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6924–6934, 2025. 2
- [25] Wufei Ma, Sky Cen, Jianzhi Shen, Rex Lee, León Begiristain, Yan Zhuang, Jiawei Peng, Zhifei Yu, Tianao Song, Xinyuan Qi, Tianmin Shu, Adam Kortylewski, and Alan Yuille. UnrealSpace: Analyzing spatial understanding and reasoning in controllable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 9716–9725, 2026. 3
- [26] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the NeurIPS Track on Datasets and Benchmarks*, 2021. 2
- [27] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling (COLM)*, 2025. 3
- [28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5389–5400, 2019. 2
- [29] Sahil Shah, S. P. Sharan, Harsh Goel, Manvik Pasula, Adithya Hebbalae, Minkyu Choi, and Sandeep P. Chinchali. Crossview: Can vision-language models reason across cameras? *arXiv preprint arXiv:2608.15539*, 2026. 3
- [30] Shanmukha Vellamcheti, Uday Kiran Kothapalli, Disharee Bhowmick, and Sathyanarayanan N. Aakur. CVT-Bench: Probing spatial-state integrity through counterfactual viewpoint transformations. *arXiv preprint arXiv:2603.21114*, 2026. Version 2, 14 August 2026. 3
- [31] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 3
- [32] Wei Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Siliang Tang, Jun Xiao, and Yueting Zhuang. Crossview suite: Harnessing cross-view spatial intelligence of MLLMs with dataset, model and benchmark. *arXiv preprint arXiv:2605.18621*, 2026. 3
- [33] Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpatialScore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31029–31041, 2026. 2
- [34] Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16789–16799, 2026. 3
- [35] Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. ST-Think: How multimodal large language models reason about 4D worlds from ego-centric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5174–5183, 2026. 3
- [36] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Matt Feiszli, and Kevin J. Liang. Multi-SpatialMLLM: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31078–31088, 2026. 3
- [37] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. *arXiv:2412.14171*. 2
- [38] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026. *arXiv:2505.23764*. 1, 3
- [39] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. MindCube: Spatial mental modeling from limited views. In *ICLR*, 2026. *arXiv:2506.21458*. 3
- [40] Leyuan Yu, Xiao Tang, Minghao Liu, Xinyuan Li, Xiaokai Bai, Sheng Zhou, Qunshu Lin, Weihao Xuan, and Naoto Yokoya. MindEdit-Bench: Benchmarking object-level counterfactual spatial reasoning in VLMs from in-the-wild photos. *arXiv preprint arXiv:2607.00491*, 2026. 3
- [41] Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *ICLR*, 2026. *arXiv:2509.25390*. 3
- [42] Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-guided modality fusion for spatial reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16811–16822, 2026. 3
- [43] Hao Zhong, Muzhi Zhu, Shenyan Zeng, Anzhou Li, Cong Chen, Hua Geng, Duochao Shi, Wentao Ye, Tao Lin, Hao

Chen, and Chunhua Shen. Eliciting complex spatial reasoning in MLLMs through wide-baseline matching, 2026. [3](#)