Reason, Then Re-reason: Cross-view Revisiting Improves Spatial Reasoning

Chaofan Ma*, Zhenjie Mao*, Yuhuan Yang, Fanqin Zeng, Yue Shi, Yingjie Zhou, Xiaofeng Cao, Jiangchao Yao

https://zhenjiemao.github.io/ReRe/

Problem Formulation

Given a first-person egocentric video of a scene and a natural-language spatial query (e.g., object counting, absolute/relative distance, object/room size, relative direction, route planning, appearance order), a Multimodal LLM must output an answer; reasoning about 3D geometry, layouts, and object relations from only what a hand-held camera happened to record.

Why Single-turn Spatial Reasoning Is Fragile

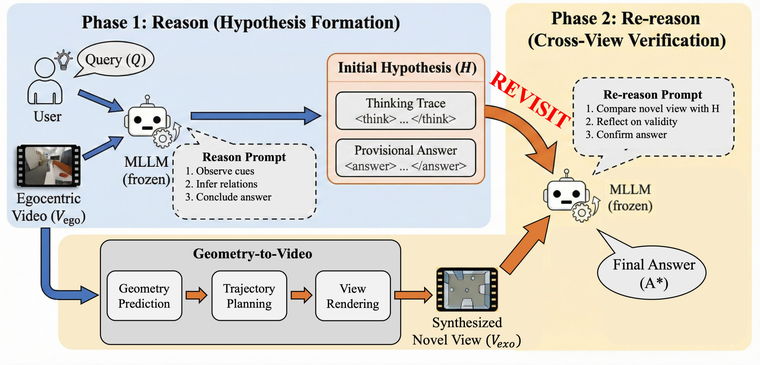
Egocentric videos are trajectory-conditioned:

- The visual evidence is limited to what the recorded camera path happens to reveal, and the temporal order of frames rarely aligns with the scene's actual spatial topology.
- Compounding this, general-purpose MLLMs lack explicit 3D world modeling and only implicitly enforce cross-frame correspondence.
- So when a model must answer in one pass, it fills missing geometry with semantic priors and produces a plausible but wrong conclusion.

Motivation

- Ideally, such errors could be detected and corrected if ***complementary cross-view evidence*** were available.
- This suggests that improving spatial reasoning is not only about better understanding in single-turn, but about enabling the model to ***revisit*** its hypothesis under complementary cross-view evidence.

Framework: Reason, Then Re-reason (ReRe)



Given an egocentric video and a spatial query, ReRe operates in two phases:

- The ***Reason Phase*** forms an initial hypothesis from the original view.
- The ***Re-reason Phase*** verifies or revises it against a synthesized allocentric view.
- A Geometry-to-Video pipeline produces this complementary view from recovered 3D geometry via trajectory planning and view rendering.

What Makes ReRe Work

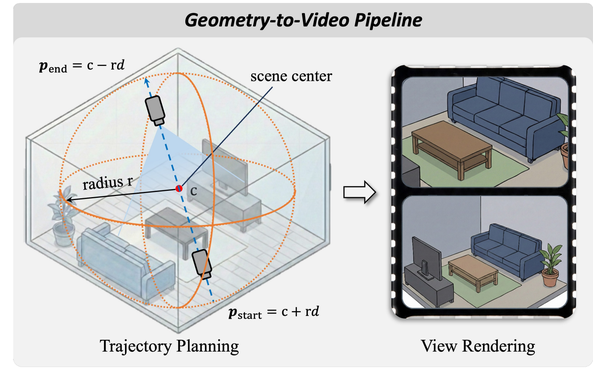
- The revisiting process is what matters.
- New evidence, not just thinking twice.
- The two views are complementary.
- Oblique Sweep is the better trajectory.

See more exciting findings and experiments results in the paper!

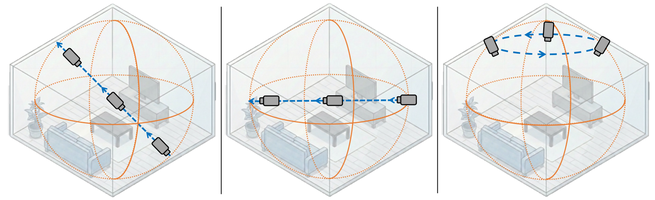
Making 3D Evidence Visible to MLLMs

The synthesized view is designed around two principles.

- ***Geometric complementarity*** requires a viewpoint that reduces inter-object occlusion and maximizes spatial coverage.
- ***Native compatibility*** means geometric evidence must present in a visual format that frozen MLLMs can handle it without architectural modifications.



ReRe predicts a 3D point cloud from the egocentric video via VGGT, plans a scene-spanning Oblique Sweep trajectory, and renders the point cloud into temporally coherent video frames via point-based rasterization.



(a) Oblique Sweep (b) Mid-level Traverse (c) Bird-eye's Orbit

Allocentric trajectory design. The Oblique Sweep follows a diagonal path through the scene center with an elevated tilt. The Mid-level Traverse moves horizontally along the diameter at a fixed elevation, while the Bird's-eye Orbit circles the scene center from a top-down perspective.

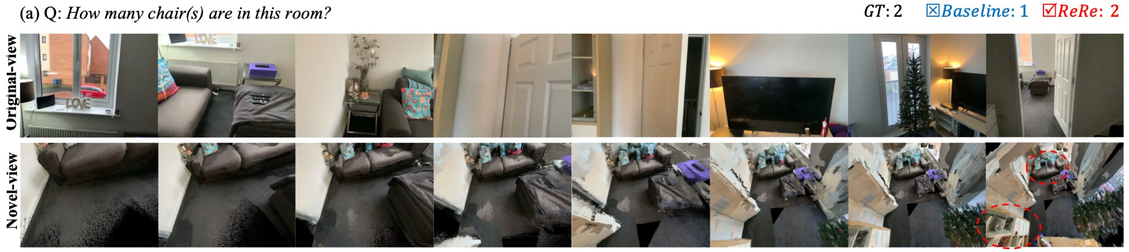
Qualitative Results

(a) Q: *How many chair(s) are in this room?*

GT: 2 ☒*Baseline*: 1 ☑*ReRe*: 2

Reason Phase output : "let me think, I need to carefully examine the video frames to identify any chairs. In the first frame, I see a bed and some items on the floor, but no chair is visible. In the second frame, I see a desk with some items on it, and there's a chair partially visible behind the desk. Wait, I should double-check if there are any other chairs in the room. Looking at the second frame again, I can see the chair more clearly—it's a standard desk chair. …… Hmm, I think that's it. There's only one chair visible in the room.\n\n<answer>1</answer>"

Re-Reason Phase output : "In the original video, I identified one chair, which was a desk chair partially visible behind the desk. Now, examining the new-view video, I can see the same desk chair more clearly. Additionally, I can now see another chair in the room — a wooden chair with a cushion, located near the bed on the left side of the frame. Therefore, there are two chairs in the room.\n\n<answer>2</answer>"



(b) Q: *You are a robot beginning at the sofa and facing the TV. You want to navigate to the bookshelf. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. Go forward until the TV 2. [please fill in] 3. Go forward until the bookshelf. You have reached the final destination. A. Turn Left, B. Turn Right, C. Turn Back*

☒*Baseline*: A ☑*ReRe*: B *GT*: B

Reason Phase output : "let me think, I'm currently facing the TV, which means I'm looking towards the front of the room. To reach the bookshelf, I need to determine its location relative to my current position. ..... Wait, I need to think about the room's layout more carefully. Since I'm facing the TV, and the bookshelf is likely to be on one side of the room, I should turn either left or right to find it. Let's break it down: if I turn left, I might see the bookshelf, or if I turn right, I might see it. I need to choose the direction that will lead me to the bookshelf. Hmm, I think I should turn left, as that's the most logical direction to find the bookshelf based on typical room layouts. Let me go with that.\n\n<answer>A</answer>"

Re-Reason Phase output : "In the original video, I assumed that turning left would lead to the bookshelf based on typical room layouts. However, in the new-view video, I can now see the actual layout more clearly. The bookshelf is located to the right of the TV, not to the left. Therefore, my previous conclusion of turning left is incorrect. The correct action is to turn right to reach the bookshelf.\n\n<answer>B</answer>"

Experiments

Performance on VSI-Bench. ReRe boosts spatial reasoning across diverse MLLMs. See more in our paper.

Methods	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
		Numerical Answer				Multiple-Choice Answer			
QWEN2.5-VL-3B	26.4	15.8	23.9	33.4	27.8	20.5	34.4	29.9	21.5
OURS	28.2$_{+1.8\%}$	16.7$_{+0.9\%}$	25.0$_{+1.1\%}$	35.3$_{+1.9\%}$	25.3$_{-2.5\%}$	31.4$_{+9.9\%}$	35.7$_{+1.3\%}$	28.9$_{-1.0\%}$	17.3$_{-4.2\%}$
QWEN2.5-VL-7B	24.8	11.2	12.2	22.8	29.9	33.8	36.0	32.0	24.9
OURS	29.5$_{+4.7\%}$	18.0$_{+6.8\%}$	18.8$_{+6.6\%}$	40.0$_{+17.2\%}$	31.1$_{+1.2\%}$	34.1$_{+0.3\%}$	35.3$_{-0.7\%}$	32.5$_{+0.5\%}$	22.0$_{-2.9\%}$
QWEN3-VL-2B	22.5	14.2	14.7	29.8	10.8	19.9	34.1	23.7	19.4
OURS	31.0$_{+8.5\%}$	17.4$_{+3.2\%}$	23.4$_{+8.7\%}$	50.5$_{+20.7\%}$	21.0$_{+10.2\%}$	25.8$_{+5.9\%}$	36.7$_{+2.6\%}$	27.8$_{+4.1\%}$	26.2$_{+6.8\%}$
QWEN3-VL-4B	30.7	14.8	22.0	41.6	33.5	30.1	38.4	23.7	29.5
OURS	36.5$_{+5.8\%}$	21.1$_{+6.3\%}$	26.7$_{+4.7\%}$	50.2$_{+8.6\%}$	35.8$_{+2.3\%}$	37.2$_{+7.0\%}$	43.1$_{+4.7\%}$	28.9$_{+5.2\%}$	34.1$_{+4.7\%}$
QWEN3-VL-8B	30.5	16.4	20.7	43.0	28.0	36.8	35.1	22.7	26.9
OURS	35.8$_{+5.2\%}$	19.5$_{+3.1\%}$	25.8$_{+5.1\%}$	49.8$_{+6.7\%}$	31.3$_{+3.3\%}$	39.7$_{+3.0\%}$	41.0$_{+5.9\%}$	29.4$_{+6.7\%}$	33.8$_{+7.0\%}$
INTERNVL2.5-4B	31.3	37.6	24.5	37.4	22.7	32.0	33.2	27.8	26.9
OURS	32.3$_{+1.0\%}$	35.8$_{-1.9\%}$	23.3$_{-1.2\%}$	36.9$_{-0.5\%}$	23.7$_{+1.1\%}$	32.5$_{+0.6\%}$	42.1$_{+8.9\%}$	30.9$_{+3.1\%}$	22.8$_{-4.0\%}$
INTERNVL2.5-8B	35.5	22.5	28.4	45.6	35.3	35.6	43.3	32.5	29.9
OURS	36.7$_{+1.2\%}$	19.1$_{-3.4\%}$	29.7$_{+1.3\%}$	46.7$_{+1.1\%}$	37.9$_{+2.6\%}$	36.5$_{+0.9\%}$	46.9$_{+3.6\%}$	31.4$_{-1.1\%}$	32.0$_{+2.1\%}$
INTERNVL3-2B	26.5	42.2	22.8	26.4	17.6	25.2	34.3	25.8	10.8
OURS	29.9$_{+3.4\%}$	37.9$_{-4.3\%}$	24.3$_{+1.5\%}$	27.2$_{+0.8\%}$	16.6$_{-1.0\%}$	32.0$_{+6.8\%}$	43.2$_{+8.9\%}$	26.8$_{+1.0\%}$	18.3$_{+7.5\%}$
INTERNVL3-8B	32.6	40.4	23.3	43.4	30.1	34.9	33.4	30.9	19.3
OURS	35.5$_{+2.9\%}$	38.8$_{-1.6\%}$	31.1$_{+7.8\%}$	44.4$_{+1.0\%}$	30.0$_{-0.1\%}$	37.2$_{+2.3\%}$	37.3$_{+3.9\%}$	30.9$_{+0.0\%}$	23.6$_{+4.4\%}$