Perception, Not Reasoning, Limits Video Spatial Understanding

Jacob Yeung[1], Mohit Goyal[2], Florian Dubost[3], Gary L. Vondran Jr.[2], Federico Tombari[2], Deva Ramanan[1], and Michael J. Tarr[1]

[1] Carnegie Mellon University
[2] Google
[3] Unity

Abstract. Wearable AI assistants must track objects across egocentric video to help users in physical spaces. Current vision-language models (VLMs) struggle with spatial questions that require combining observations from different moments. An overall score does not reveal whether their errors come from perceiving the scene or reasoning about it. We introduce SPLIT, a training-free harness that reconstructs one metric 3-D point cloud from a video and gives a planner VLM tools to query it. Frozen reconstruction and grounding models locate objects and measure their geometry, and the planner VLM combines those measurements using Python. We train no model parameters on the evaluated benchmarks. SPLIT improves accuracy on VSIBench, VSTIBench, and ReVSI. On VSTIBench, it raises accuracy from 57.7% to 78.3%, eliminating 49% of the baseline VLM's errors; across all three benchmarks, the largest gains occur when questions relate objects seen in different frames. SPLIT achieves the best published training-free score on VSIBench and exceeds published scores on VSTIBench and ReVSI. To measure how much error comes from perception, we rerun SPLIT with ground-truth measurements in place of the tools' estimates. On the VSTIBench analysis subset, the baseline VLM scores 59.0%, SPLIT scores 79.7%, and SPLIT + GT reaches 96.7%. This gain shows that inaccurate perception remains a major source of error.

Keywords: video spatial understanding · agentic harness

1 Introduction

Wearable assistants must answer questions about objects observed at different moments: where did I leave my keys, will this couch fit along that wall, or how do I return to the kitchen? Answering these questions requires localizing objects, recovering camera motion and metric geometry, and registering observations from different frames in one coordinate system. VSIBench [16], VSTIBench [4], and ReVSI [20] test metric quantities, such as room dimensions and inter-object distances, and qualitative relations, such as directions, order, and routes. Yet state-of-the-art vision-language models (VLMs) remain well below human accuracy on these tasks. An end-to-end score obscures two abilities:

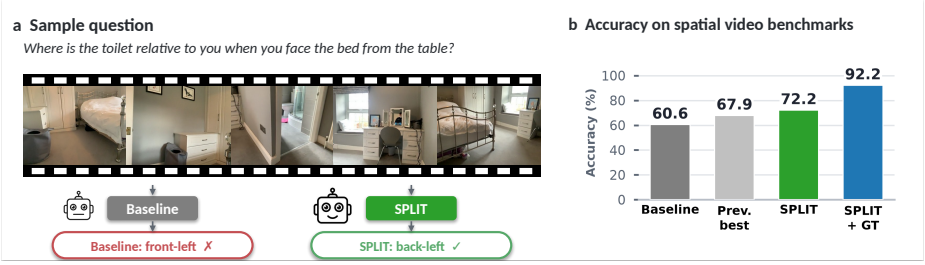

Fig. 1: Video Spatial QA with and without Perception Tools. Perception tools help the planner VLM relate objects observed in different frames. (a) The named objects are not covisible, so no single frame contains their relation. The baseline VLM incorrectly answers *front-left*. SPLIT grounds the named objects in one metric 3-D space and answers *back-left* correctly. (b) Averaged over the full VSIBench, VSTIBench, and ReVSI evaluations, the baseline VLM scores 60.6% and SPLIT scores 72.2%. SPLIT + GT scores 92.2% when averaged over the three benchmark-specific analysis subsets; VSIBench and ReVSI counting are oracle ceilings. The previous best averages the best published score on each benchmark (a different model on each, including trained specialists).

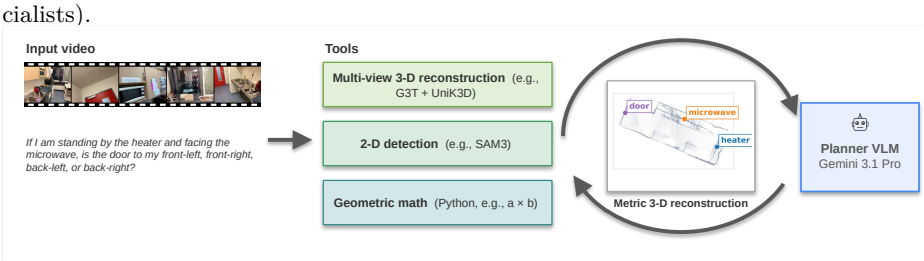

Fig. 2: SPLIT Pipeline Overview. SPLIT provides the planner VLM with metric measurements registered in one scene frame. Multi-view reconstruction provides camera poses and a metric point cloud, while SAM3 or a grounding VLM locates queried objects in 2-D, and generated code combines the returned geometry. The planner VLM may call these perception tools repeatedly before answering.

- **Perception** converts pixels into object regions, camera poses, depth, and metric object geometry. Segmenters, detectors, and grounding VLMs can all provide object regions.
- **Reasoning** composes those measurements into directions, distances, counts, sizes, relations, and routes.

For example, when a model answers "2.4 m" incorrectly, its score cannot distinguish a mislocalized object from an incorrectly computed distance.

We implement this separation with SPLIT (Separating Perception from LLM Inference with Tools), a training-free harness for spatial video. The reconstruction and grounding models and the planner VLM remain frozen, and we make no benchmark-specific parameter updates. Gemini 3.1 Pro serves as the planner VLM. It finds relevant frames, grounds queried objects in 2-D, reads their 3-D coordinates from a shared metric point cloud, queries camera poses, and writes code to combine the measurements. We evaluate SPLIT on VSIBench,

VSTIBench, and ReVSI. On VSIBench, SPLIT scores 66.9% under the official metric, the best published training-free result to our knowledge, but below the human reference of 79.2% (Section 5).

We replace the perception values behind the fixed tool interface while keeping the planner VLM, prompt, and questions unchanged. On the VSTIBench analysis subset, the baseline VLM scores 59.0%, SPLIT scores 79.7%, and SPLIT + GT reaches 96.7% (Section 5). This 16.9% gain shows that accurate perception recovers substantial headroom. The traces link the category regressions to planner VLM answers that ignore better measurements and to masks that miss or overshoot objects.

Contributions.

1. **SPLIT improves spatial video QA.**
   SPLIT reduces the baseline VLM's remaining error by a factor of up to 1.9× across three benchmarks, with the largest gains on questions that relate objects seen in different frames. We will release the harness and code.
2. **Ground-truth perception improves performance.**
   On the VSTIBench analysis subset, SPLIT + GT reaches 96.7% versus 79.7% for SPLIT and 59.0% for the baseline VLM.
3. **Bad masks and evidence selection explain the measured size regressions.**
   Where the perception tools lose to the baseline VLM, the traces show two distinct failures: masks cover only part of the object or spill past it, or the planner VLM answers with a worse value than one it already measured. These failures dominate the measured regressions; global metric-scale error is not the primary cause.

2 Related Work

Supervised spatial specialists. Supervised spatial specialists fine-tune VLMs on spatial training data. Recent examples include VLM-3R [4], Cambrian-S [18], SpaceMind++ [5], Cambrian-P [17], SSR-3D [19], and Think-with-Spatial-Code [3]. Tables 1 and 2 compare their published scores with our training-free method.

Training-free methods. Tool-using VLMs call detectors, depth models, or code interpreters to improve compositional tasks [6, 15]. On spatial understanding, TRACE prompts the VLM itself to write a structured scene description as intermediate reasoning, with no external perception [7]. ViSRA reconstructs a shared, gravity-aligned scene frame at relative scale [11]. pySpatial generates Python visual programs over 3-D tools such as reconstruction, pose recovery, and novel-view rendering [10]. RieMind lets an LLM agent query geometric tools over a persistent 3-D scene graph built from ground-truth annotations [14]. These systems answer from a prepared scene representation or a one-shot generated program. SPLIT constructs a metric point cloud with camera poses from the video and lets the planner VLM query this representation as it works. The fixed interface lets us measure how much the reconstruction helps and how much error remains with ground-truth perception.

Table 1: Video Spatial Reasoning Benchmark Results. SPLIT improves performance across video spatial-reasoning benchmarks. Bold and underlining mark the best and second-best full-benchmark scores. †Ground-truth cells use the corresponding analysis subsets described in Section 4; VSIBench and ReVSI counting are oracle ceilings. Tables 2, 3, and 4 give per-category scores.

Method	VSIBench	VSTIBench	ReVSI
Proprietary			
GPT-4o	34.0	38.2	—
Gemini-2.5 Pro	51.5	42.4	—
GPT-5.2	49.2	—	50.9
Gemini 3 Flash	55.9	—	57.6
Gemini 3 Pro	60.5	—	60.9
Open-source			
Qwen3-VL-8B	57.4	—	49.1
InternVL3.5-38B	60.8	—	54.1
LLaVA-Video-72B	39.7	44.0	37.8
Supervised spatial specialists			
VLM-3R-7B [4]	60.9	58.8	50.1
Cambrian-S-7B [18]	67.5	—	49.1
GeoThinker-8B	72.6	67.4	—
Cambrian-P [17]	<u>73.7</u>	<u>68.9</u>	—
SSR-3D [19]	**73.9**	44.8	—
Baseline VLM (Gemini 3.1 Pro)	60.4	57.7	<u>63.8</u>
SPLIT w/ Gemini 3.1 Pro (Ours)	66.9 (+6.5)	**78.3** (+20.6)	**71.5** (+7.7)
SPLIT (Ours) + GT	88.8 †	96.7 †	91.0 †

Perception tools and composition. The planner VLM composes measurements returned by task-general perception tools. The tools receive object names from the question but encode no benchmark category or answer rule. 3-D scene reconstruction provides a shared metric world frame, open-vocabulary grounding provides object regions, and metric depth supplies physical scale [8, 12, 13].

3 SPLIT: A Training-Free Harness

Every SPLIT perception tool returns a measurement, such as frames, masks, 3-D points, or camera poses. The planner VLM composes these measurements into the answer (Figure 2).

3.1 World Frame from Video

Objects named in a question rarely share a frame, so SPLIT first registers the camera's observations as one metric point cloud with camera poses. We reconstruct each scene once and reuse the cached result for every question about that scene. For each selected frame i, a feed-forward reconstruction model predicts a camera-to-world pose $\mathbf{T}_i \in \mathrm{SE}(3)$ and dense per-frame 3-D points $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ with confidence and validity masks. The reconstructed points and poses share a gravity-aligned, Z-up world coordinate system, so object locations, extents, and camera motion can be compared across views. We use one reconstruction for every category: G3T predicts the per-frame 3-D points, and UniK3D supplies one metric scale per scene [13].

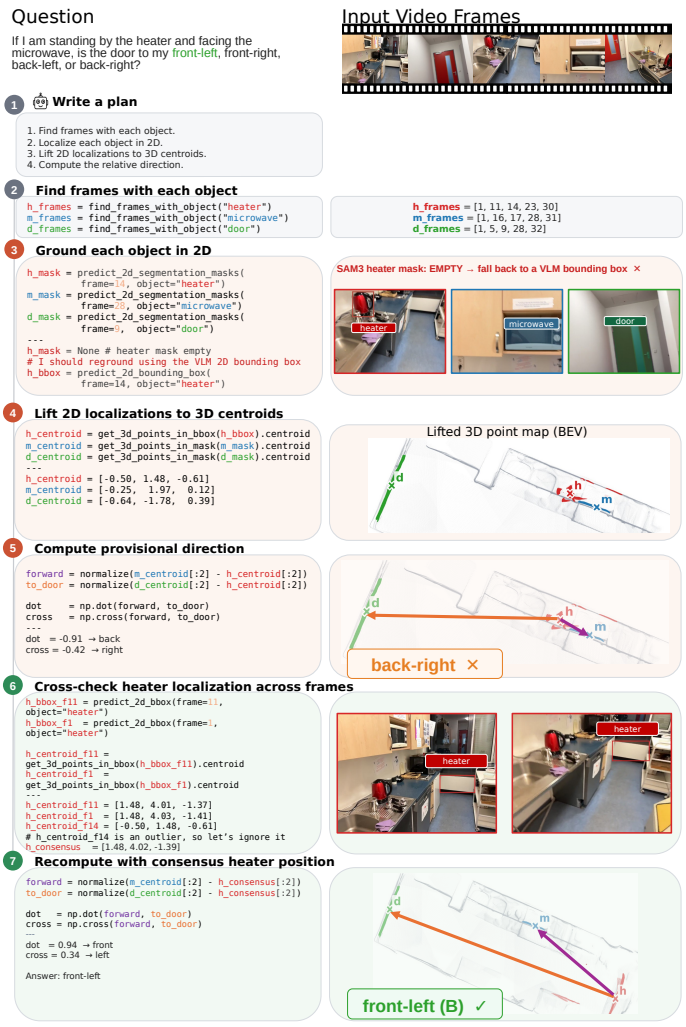

Fig. 3: Example SPLIT Trace. SPLIT rechecks inconsistent grounding before answering. One recorded episode is shown in Steps 1–7, with planner outputs on the left and execution outputs on the right. Initially, SAM3 returns no mask for the heater, while the fallback VLM box lands on a kettle (3, red), producing the provisional direction *back-right* (4–5). The planner VLM then grounds the heater in two more frames; those measurements agree and expose the first box as an outlier (6). After discarding the outlier, the planner VLM recomputes the correct direction, *front-left* (7).

3.2 Perception Tools

The perception tools return measurements rather than answers. We call an operation *perception* when it converts pixels into an object region or 3-D geometry, whether SAM3 or a grounding VLM produces the region. The **2-D perception** tools retrieve frames likely to show a named object and ground the object as a box, points, or a SAM3 mask. The **3-D measurement** tools lift a grounded region through $\mathbf{P}_i$ into world coordinates, aggregate those points into positions and extents, and return camera poses. We also let the planner VLM execute Python over the returned measurements, for example to compute distances or use a cross product to determine left-right relations.

3.3 Planner VLM Composition

The planner VLM selects perception tools and receives their logged outputs. LangGraph implements this control flow [9]. The planner VLM either writes code for the deterministic executor or combines tool-computed values directly. It calculates distances, resolves turns, estimates object extent, and deduplicates instances across views. The prompt defines coordinate and direction conventions. When measurements disagree, the planner VLM can gather more evidence before it commits to an answer. In Figure 3, an initial grounding lands on the wrong object, but two consistent measurements from other frames expose that estimate as an outlier and change the final direction.

3.4 Ground-Truth Perception

Across all three benchmarks, we replace the perception tools' estimates with the ground-truth object regions, geometry, and camera information available in each benchmark. The planner VLM still chooses tools and composes their returned values into an answer. We evaluate only questions covered by the benchmark annotations and label counting results that directly expose instance cardinality as oracle ceilings.

4 Experimental Setup

Benchmarks and metrics. We evaluate on VSIBench [16] across all eight spatial categories: object counting, absolute and relative distance, object and room size estimation, relative direction, route planning, and appearance order. We also evaluate the official *Debiased* split, which removes questions that admit non-visual shortcuts [1]. ReVSI tests questions tied to official frame samples, where a referenced object may be absent from the sampled frames [20]. VSTIBench additionally tests camera motion [4].

Analysis subsets. The VSIBench error analysis uses a category-balanced subset drawn from the published *Debiased* split [1]. The VSIBench ground-truth intervention uses a 500-question subset with 50 questions per category; every arm answers the same questions. The VSTIBench ground-truth intervention uses a 450-question subset with 50 questions per category. The ReVSI ground-truth diagnostic uses a 981-question analysis subset that excludes 18 route questions without complete annotations; its counting cell is an oracle ceiling.

Baseline VLM. The baseline VLM runs `gemini-3.1-pro` end to end without perception tools. It receives each question and its fixed 32 sampled frames with the standard answer-format prompt, and we preserve full per-question traces.

Models. We use all models off the shelf and do not train them on the evaluated benchmarks. The baseline VLM and the planner VLM inside SPLIT use the `gemini-3.1-pro` checkpoint. SAM3 segmentation [2] and `gemini-3.5-flash` provide object grounding. Camera poses and dense metric 3-D points come from the feed-forward reconstruction of Section 3.1.

Configurations. We compare: (i) the **baseline VLM**, which answers directly from sampled frames; (ii) **SPLIT**, which uses reconstructed perception; and (iii) **SPLIT + GT**, which reads ground-truth values through the perception-tool interface.

Efficiency measurements. We derive planner VLM turns and token use from the saved traces.

5 Results and Analysis

Finding 1: Perception tools help most when questions span views

Gains across benchmarks. SPLIT raises accuracy over the baseline VLM from 60.4% to 66.9% on VSIBench, from 57.7% to 78.3% on VSTIBench, and from 63.8% to 71.5% on ReVSI (Table 1). Averaged over the three benchmarks, SPLIT raises accuracy from 60.6% to 72.2%. On VSIBench, SPLIT is the best published training-free method to our knowledge, although the strongest supervised specialists remain ahead (Table 2). On VSIBench and ReVSI, the largest improvements occur in relative direction (+27.6% and +35.3%) and route planning (+15.0% and +24.4%), tasks that often require combining observations across frames. On VSTIBench, the largest improvements occur in camera-motion direction (+51.4%) and displacement (+41.7%), where camera poses and metric geometry directly supply the missing evidence. The VSIBench gain grows from 6.5% on the full benchmark (66.9% versus 60.4%) to 8.3% on the shortcut-pruned subset (65.8% versus 57.5%), so the gain persists after shortcut pruning. The gain also survives planner VLM swaps: with Opus 4.8 or GPT-5.5 as the planner VLM, SPLIT scores 61.8% and 62.1% on the 500-question analysis subset, versus 54.9% for the baseline VLM on the matching pre-repair appearance cohort (Table 5).

Finding 2: Ground-truth perception further improves performance

Ground-truth intervention. On the VSTIBench analysis subset, the baseline VLM scores 59.0%, SPLIT scores 79.7%, and SPLIT + GT reaches 96.7% (Table 3). Only the values returned by the perception tools change. Ground-truth perception therefore adds 16.9% over SPLIT while the planner VLM still composes every answer from the returned geometry.

Table 2: VSIBench Results Comparison. Among training-free methods, SPLIT has the best published overall score, exceeding the prior best, TRACE, by +9.1% under the official protocol. "TF" is training-free, "RL" reinforcement learning, and "SFT" supervised fine-tuning. Human and general-purpose results are taken from VSI-Bench and later comparison tables [16, 19]. Each Δ row is SPLIT minus the baseline VLM. Bold and underlining mark the best and second-best full-benchmark model scores. †Subset result reported by ViSRA. ‡Ground-truth diagnostic; the analysis subset is described in Section 4, counting is an oracle ceiling, and appearance order uses a fixed 50-question cohort shared by every arm.

			Numerical				Multiple choice			
Method	**Train**	**Avg.**	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Human [16]	—	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
General-purpose VLMs										
GPT-4o	TF	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Pro	TF	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Gemini-2.5 Pro	TF	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
Supervised spatial specialists										
Think-with-Spatial-Code [3]	RL	54.9	58.3	39.0	73.0	52.4	57.8	55.9	38.7	63.9
+ 2D box	RL	60.1	**95.2**	60.7	50.8	33.1	**87.1**	62.0	32.5	59.0
VLM-3R-7B [4]	SFT	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
Cambrian-S-7B [18]	SFT	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
SpaceMind++ (InternVL3-8B) [5]	SFT	73.4	74.5	62.4	**77.9**	76.9	73.5	89.7	48.2	84.1
Cambrian-P (Qwen2.5-7B) [17]	SFT	73.7	74.9	60.1	76.0	76.9	74.8	89.5	52.6	**85.0**
SSR-3D (openPangu-VL-7B) [19]	SFT	**73.9**	70.5	**67.1**	76.0	**79.5**	71.3	**93.4**	48.5	**85.0**
Training-free methods										
ViSRA (Qwen2.5-VL-7B)† [11]	TF	50.3	46.4	27.8	60.1	47.7	51.3	71.7	43.8	53.7
TRACE (Gemini 3 Pro) [7]	TF	57.8	47.6	38.8	73.9	45.6	63.9	61.7	58.0	73.0
Full benchmark										
Baseline VLM	TF	60.4	51.2	43.6	73.3	53.1	69.2	62.4	58.8	71.7
SPLIT (Ours)	TF	66.9	50.0	58.3	66.1	56.3	69.9	90.0	**73.7**	70.7
Δ		+6.5	-1.2	+14.7	-7.2	+3.2	+0.7	+27.6	+15.0	-1.0
Debiased subset										
Baseline VLM	TF	57.5	46.4	41.2	62.9	54.6	69.0	62.3	51.8	71.7
SPLIT (Ours)	TF	65.8	50.9	56.4	60.2	57.5	71.6	86.7	71.1	72.3
Δ		+8.3	+4.5	+15.2	-2.7	+2.8	+2.6	+24.4	+19.3	+0.6
Ground-truth diagnostic‡										
SPLIT (Ours) + GT	TF	88.8	93.4	76.6	91.2	92.4	88.0	98.7	70.0	100.0

Gains vary by task. Ground truth raises camera-object absolute distance from 63.8% to 89.8% and camera displacement from 56.2% to 97.6%. Camera-motion direction and relative position are already near ceiling, and camera-object relative distance rises from 85.3% to 98.0%.

Table 3: VSTIBench Results Comparison. On VSTIBench, SPLIT gains +20.6% over the baseline VLM. The Δ row is SPLIT minus the baseline VLM. Bold and underlining mark the best and second-best full-benchmark scores. *Published comparison rows use original-release labels. We identified a camera-center bug in camera displacement, camera–object distance (absolute and relative) labels and notified the VSTIBench authors; they have corrected the release, public erratum. Our rows use the corrected release labels. †Ground-truth diagnostic; the analysis subset is described in Section 4.

Method	Train	Avg.	Cam-Obj Abs (MRA)	Cam Disp (MRA)	Cam Mov Dir	Obj-Obj Rel Pos	Cam-Obj Rel Dist
*Published comparisons**							
VLM-3R-7B [4]	SFT	58.8	39.4	39.6	60.6	86.5	68.6
GeoThinker-8B [17]	SFT	67.4	38.4	45.8	84.2	<u>93.6</u>	75.2
Cambrian-P [17]	SFT	<u>68.9</u>	42.5	<u>46.6</u>	<u>87.7</u>	**94.3**	73.2
Baseline VLM	TF	57.7	<u>49.0</u>	18.5	45.3	93.5	<u>82.0</u>
SPLIT (Ours)	TF	**78.3**	**58.2**	**60.2**	**96.7**	90.4	**85.7**
Δ		+20.6	+9.3	+41.7	+51.4	-3.2	+3.7
Ground-truth diagnostic†							
Baseline VLM	TF	59.0	55.0	20.6	44.0	98.7	76.7
SPLIT (Ours)	TF	79.7	63.8	56.2	96.0	97.3	85.3
SPLIT (Ours) + GT	TF	96.7	89.8	97.6	98.0	100.0	98.0

Table 4: ReVSI Results Comparison. SPLIT gains most on relative direction and route planning. Published comparison scores are quoted from ReVSI [20]; frame counts are listed in the table. The Δ row is SPLIT minus the baseline VLM. Bold and underlining mark the best and second-best full-benchmark scores. †Ground-truth diagnostic; the analysis subset is described in Section 4, and counting is an oracle ceiling.

				Numerical				Multiple choice		
Method	Train	Frames	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
GPT-5.2	TF	64	50.9	56.2	41.5	<u>73.9</u>	**63.0**	48.4	34.9	38.2
InternVL3.5-38B	TF	64	54.1	43.8	<u>60.6</u>	70.2	58.4	57.4	45.9	42.7
VLM-3R-7B [4]	SFT	32	50.1	41.6	61.6	64.8	52.5	46.5	49.5	34.1
Baseline VLM	TF	32	<u>63.8</u>	**64.6**	59.6	**76.8**	<u>61.9</u>	<u>67.0</u>	<u>58.0</u>	<u>58.5</u>
SPLIT (Ours)	TF	32	**71.5**	<u>58.8</u>	**66.8**	64.2	50.3	**84.2**	**93.2**	**82.9**
Δ			+7.7	-5.8	+7.2	-12.5	-11.6	+17.2	+35.3	+24.4
Ground-truth diagnostic†										
SPLIT (Ours) + GT	TF	32	91.0	86.7	98.2	85.0	79.7	97.9	98.3	91.1

Finding 3: Bad masks and evidence selection explain the measured regressions

Why some categories regress. The perception tools do not improve every VSI-Bench category: object counting and appearance order remain within 1.2% of the baseline VLM, while object size drops by 7.2%. On ReVSI, counting, object size,

Table 5: Planner Swaps on the VSIBench Analysis Subset. The harness gain transfers across planner VLMs. Each row runs SPLIT with a different planner VLM on the 500-question analysis subset; the baseline VLM scores 56.9. The GPT-5.5 and Opus 4.8 evaluations predate the appearance-cohort repair (Section 4); their comparable baseline is 54.9.

Planner VLM	Avg.
Gemini 3.1 Pro	**65.3**
GPT-5.5	62.1
Opus 4.8	61.8

and room size drop by 5.8%, 12.5%, and 11.6%. Across VSIBench objects with ground-truth correspondences, the smallest axis, which usually points in depth, is 25–42% too thin, while the longest axis used for object size is only 9% too short. For 80% of VSIBench objects, at least one tool measurement is accurate enough to score; when SPLIT still loses on these, the planner VLM answered with a different number than that measurement. For the remaining 20%, no tool measurement is accurate enough to score, mostly because the masks cover only part of the object. On ReVSI, only 7.5% of object-size answers lack a 3-D size measurement. In 224 of the 429 questions where the perception tools lose to the baseline VLM, the planner VLM had already measured a size that would score higher than its final answer. Measurement errors from masks that cover only part of the object or spill past it account for about a third of the gross loss. These are distinct failures: bad masks are perception errors, while discarding a better measurement is a planner VLM evidence-selection error. On the ReVSI ground-truth diagnostic described in Section 4, SPLIT + GT reaches 91.0%; its counting score is an oracle ceiling. Counting loses in three comparable ways: the grounding tools return no usable instance count, the planner VLM ignores a correct count a tool returned, and it double-counts or misses instances across frames.

6 Limitations

SPLIT requires substantial inference-time computation: on the VSIBench analysis subset, it calls the planner VLM a median of 16 times and generates a median of 5,980 tokens, including thinking tokens, per question. All three benchmarks derive from static indoor scans, so dynamic egocentric video remains untested. The ground-truth interventions replace several perception components together, so they measure combined perceptual headroom rather than isolate one component.

7 Conclusion

Spatial video question answering requires recovering a scene from pixels and composing its geometry into an answer. SPLIT separates these stages: frozen perception models supply measurements, and a planner VLM composes them. Across three benchmarks, SPLIT reduces the baseline VLM's remaining error

by a factor of up to 1.9×. With ground-truth perception, the same planner VLM reaches 96.7% on VSTIBench, 88.8% on VSIBench, and 91.0% on ReVSI. These results show that frontier planner VLMs can compose accurate spatial values and that better perception can unlock substantially more of this capability. SPLIT therefore provides a training-free way to measure how much spatial reasoning across an entire video improves with more accurate perception.

References

1. Brown, E., Yang, J., Yang, S., Fergus, R., Xie, S.: Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. arXiv preprint arXiv:2511.04655 (2025)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Ravi, N., Dollár, P., Feichtenhofer, C., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Chen, J., Ma, W., Yuan, R., Zhang, Y., Wu, J., Yuille, A.: Thinking with spatial code for physical-world video reasoning. arXiv preprint arXiv:2603.05591 (2026)
4. Fan, Z., et al.: VLM-3R: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
5. Gu, B., Zhang, Z., Wei, Z., Chen, Z., Li, L., Song, Z.: Spacemind++: Toward allocentric cognitive maps for spatially grounded video mllms. arXiv preprint arXiv:2605.09449 (2026)
6. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14953–14962 (2023)
7. Hua, J., Yin, Y., Wu, Y., Wang, T., Huang, Y., Liu, M.: Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. arXiv preprint arXiv:2603.23404 (2026)
8. Keetha, N., Karaev, N., Engel, N., Lin, X., Zhou, X., Wang, Y., Aigerman, N., Sucar, E., Tagliasacchi, A., Held, D., Hinterstoisser, S., Ramanan, D., Novotny, D., Iyer, K.M.: MapAnything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
9. LangChain, Inc.: LangGraph. https://github.com/langchain-ai/langgraph (2024)
10. Luo, Z., Zhang, C., Yong, S., Dai, C., Wang, Q., Ran, H., Shi, G., Sycara, K., Xie, Y.: pyspatial: Generating 3d visual programs for zero-shot spatial reasoning (2026)
11. Mou, T., He, J., Wang, R., Liu, C., Yang, H., Zhang, T., Chen, J., Ma, X.: ViSRA: A video-based spatial reasoning agent for multi-modal large language models. arXiv preprint arXiv:2605.10106 (2026)
12. Nagoor Kani, B.R., Snavely, N.: G3T up! gravity aligned coordinate frames simplify pointmap processing. arXiv preprint arXiv:2605.27372 (2026)
13. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: UniK3D: Universal camera monocular 3d estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
14. Ropero, F., Turkoz, E., Matos, D., Du, J., Ruiz, A., Zhang, Y., Liu, L., Sun, M., Wang, Y.: Riemind: Geometry-grounded spatial agent for scene understanding (2026)
15. Surís, D., Menon, S., Vondrick, C.: ViperGPT: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11888–11898 (2023)
16. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. arXiv preprint arXiv:2412.14171 (2024)
17. Yang, J., Zhao, Z., Pan, X., Yang, S., Zhang, J., Kang, B., Xu, H., Xie, S.: Cambrian-P: Pose-grounded video understanding. arXiv preprint arXiv:2605.22819 (2026)

18. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Cambrian-S: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)
19. Zhang, Y., Xia, Y., Wang, Y., Song, M., Wu, X., Wan, W., Liu, B., Ye, A., Zhang, H., Wen, F.: SSR: Pushing the limit of spatial intelligence with structured scene reasoning. arXiv preprint arXiv:2603.00409 (2026)
20. Zhang, Y., Chen, J., Tan, J., Mao, Y., Chen, W., Chang, A.X.: ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3D reasoning. arXiv preprint arXiv:2604.24300 (2026)