PERCEPTION, NOT REASONING, LIMITS VIDEO SPATIAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Current vision-language models (VLMs) struggle with spatial reasoning across time, and aggregate benchmark scores obscure whether failures arise from perception or reasoning. We introduce SPLIT (Separating Perception from LLM Inference with Tools), a training-free harness for video spatial reasoning that builds a metric 3-D reconstruction of a video and provides a base VLM with tools to query it. SPLIT improves accuracy on four video benchmarks (VSIBench, VSTIBench, ReVSI, and DSI-Bench) and the multi-image MMSI-Bench by up to 20.6% over the base VLM answering without tools, with gains extending to open-weight base VLMs. SPLIT helps more when a question's objects are not co-visible in any frame: on VSIBench relation questions, its improvement over the base VLM alone nearly doubles when no frame shows all the objects a question names. Accurate perception closes most of the gap to a perfect score: when the tools return ground truth (SPLIT + GT), that gap shrinks 3.6-fold on subsets of VSIBench, VSTIBench, and ReVSI, and accuracy reaches an average of 92.9%; two open-weight base VLMs nearly match it, with Gemma-4-31B averaging 90.4% and Qwen3.6-27B 88.7%. We will release the harness and code.

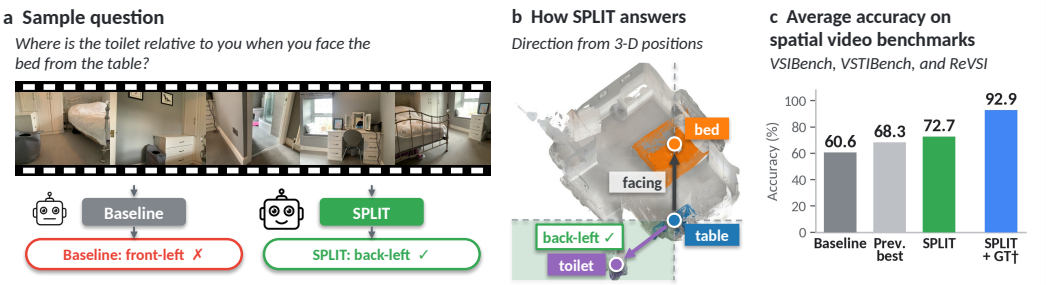


Figure 1: **Perception Tools for Video Spatial Questions.** SPLIT allows the base VLM to relate objects across frames by composing perception tools. (a) The three objects in the question never appear in the same frame. Without tools, the base VLM answers *front-left*, which is wrong. (b) SPLIT locates all three objects in one 3-D reconstruction of the video and computes the direction from their positions. Appendix Figure 6 shows the full trace. (c) We show the average accuracy across VSIBench, VSTIBench, and ReVSI. †We evaluate SPLIT + GT, whose tools return ground-truth measurements, on a subset of each benchmark (Section 4).

1 INTRODUCTION

Questions about physical space often combine observations made at different times: Where did I leave my keys? Will this couch make it up the staircase? How do I return to the kitchen? To answer these questions, a model must localize objects, recover camera motion and geometry, and align observations across frames in one world coordinate system. VSIBench (Yang et al., 2024), VSTIBench (Fan et al., 2025), and ReVSI (Zhang et al., 2026) test both quantities, such as room dimensions and distances between objects, and relations, such as directions, order, and routes. Yet state-of-the-art

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.

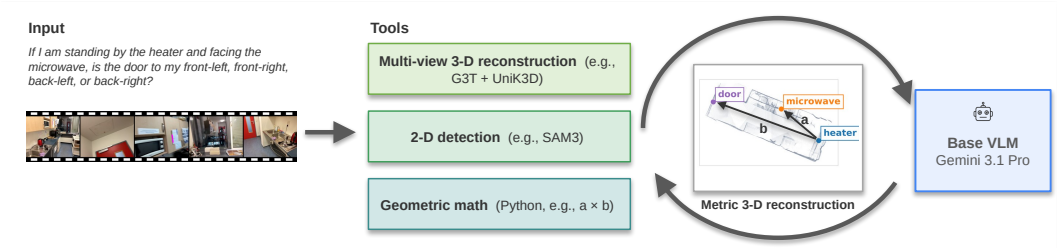


Figure 2: **SPLIT Overview.** A base VLM answers a spatial question by composing perception tools that measure the scene in one world coordinate system. Arrows **a** and **b** run from the heater to the microwave and to the door. Appendix Figure 7 shows the full trace of this example.

vision-language models (VLMs) still score well below humans on these tasks. End-to-end scores conflate two distinct abilities:

- **Perception**: extracting object regions, camera poses, depth, and object geometry from pixels, whether a segmenter, a detector, or a grounding VLM supplies the regions.
- **Reasoning**: composing these measurements to infer directions, distances, counts, sizes, relations, and routes.

For example, an incorrect answer of "2.4 m" does not reveal whether the model mislocalized an object or miscomputed the distance.

We implement this separation in SPLIT (**S**eparating **P**erception from **LLM** **I**nference with **T**ools), a training-free harness for video spatial reasoning. The base VLM, Gemini 3.1 Pro, composes tools that find relevant frames, ground queried objects in 2-D, read their 3-D coordinates from one 3-D reconstruction, and return camera poses. It then writes code that combines these measurements. To isolate what perception contributes, we let the tools return ground truth on the same questions, and on VSIBench we also give the base VLM each question's task and object labels (Section 3.4).

Over the base VLM answering without tools, SPLIT raises the score on all five benchmarks, by 7.5% to 20.6% (Table 1), and in nine of ten comparisons with the open-weight base VLMs (Table 2). When the tools return ground truth instead, SPLIT improves from 68.5% to 91.9% on a VSIBench subset and from 81.5% to 95.7% on a VSTIBench subset (Figure 4).

The task-level results reveal a boundary: geometry helps directions and routes more consistently than object extents (Figure 4).

Contributions.

1. **A training-free harness for video spatial reasoning.** SPLIT improves accuracy on four video benchmarks (VSIBench, VSTIBench, ReVSI, and DSI-Bench) and the multi-image MMSI-Bench (Table 1). We will release the harness and code.
2. **Accurate perception closes most of the gap to a perfect score.** When the tools return ground truth (SPLIT + GT), the gap to a perfect score shrinks 3.6-fold on subsets of VSIBench, VSTIBench, and ReVSI, and accuracy reaches an average of 92.9% (Figure 4). Two open-weight base VLMs nearly match it: Gemma-4-31B averages 90.4% and Qwen3.6-27B 88.7% (Table 2).
3. **SPLIT helps more when a question's objects are not co-visible in any frame.** On VSIBench relation questions, SPLIT's improvement over the base VLM alone nearly doubles when no frame shows all the objects a question names (Appendix Table 8).

2 Related Work

Supervised spatial specialists. These methods fine-tune VLMs on spatial training data. Recent examples include VLM-3R (Fan et al., 2025), Cambrian-S (Yang et al., 2025a), and Cambrian-P (Yang et al., 2026). Table 1 and Appendix Table 3 compare their published scores with ours.



Figure 3: **Example SPLIT Plans.** SPLIT composes different perception tools for different spatial questions about one video. Each column shows a question, a summary of the base VLM's tool calls, a 3-D view of the queried objects or cameras, and the answers of the base VLM, SPLIT, and the ground truth. Appendix Figures 8, 9, and 10 show the full traces.

Training-free methods. ReAct interleaves reasoning steps with tool calls (Yao et al., 2023), and tool-using VLMs call detectors, depth models, or code interpreters to solve compositional tasks (Surís et al., 2023; Gupta & Kembhavi, 2023). For spatial understanding, TRACE prompts the VLM to write a structured scene description as intermediate reasoning, without external perception (Hua et al., 2026). SPLIT uses external perception models to build a metric 3-D reconstruction with camera poses from the video alone and lets the base VLM query it repeatedly while reasoning. Concurrent work, SpatialClaw (Cho et al., 2026), also equips a VLM agent with perception and geometry tools, which the agent calls by writing Python code in a persistent notebook. SpatialClaw studies how to design this code interface, while SPLIT measures how much of the remaining error comes from perception: its tools can return ground-truth values in place of estimates, so we can rerun the same agent with perfect perception. Table 1 compares SPLIT with SpatialClaw's published results, which use Qwen3.6-27B as the base VLM.

Perception models as tools. Feed-forward 3-D reconstruction provides a world coordinate system (Wang et al., 2024; 2025; Keetha et al., 2025; Nagoor Kani & Snavely, 2026), open-vocabulary segmentation provides object regions (Carion et al., 2025), and monocular metric depth supplies physical scale (Piccinelli et al., 2025). SPLIT uses such models as perception tools whose measurements the base VLM composes. The tools receive object names from the question but encode no benchmark task or answer rule.

3 SPLIT: A Training-Free Harness

Every SPLIT perception tool returns an intermediate result, such as frames, masks, 3-D points, or camera poses. The base VLM composes these results into an answer (Figure 2).

3.1 Metric 3-D Reconstruction from Video

Objects named in a question rarely share a frame, so SPLIT first registers the camera's observations as one metric 3-D reconstruction with camera poses. We reconstruct each scene once and reuse the cached result for every question about that scene. For each selected frame i, the feed-forward reconstruction model G3T (Nagoor Kani & Snavely, 2026) 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. Because G3T predicts gravity-aligned geometry, the reconstructed points and poses share a Z-up world coordinate system, so the base VLM can compare object locations, extents, and camera motion across views. UniK3D (Piccinelli et al., 2025) sets each scene's metric scale to the median over frames of the ratio between its median depth and G3T's, and SPLIT uses this reconstruction for every task.

3.2 Perception Tools

We call an operation *perception* when it converts pixels into an object region or 3-D geometry. Some perception tools retrieve frames likely to show a named object and ground the object as a SAM3 mask (Carion et al., 2025) or as a box or points from the `gemini-3.5-flash` grounding VLM; the prompt makes the box the fallback when SAM3 finds no instance. Others lift a grounded region through $\mathbf{P}_i$ into world coordinates, aggregate those points into positions and extents, and return camera poses. The base VLM can also run Python on the returned measurements, for example to compute distances or to decide left from right with a cross product. Appendix Table 9 lists the perception tools and the Python executor.

3.3 Composing Measurements

The base VLM works in a loop: it selects a perception tool, reads the output, and decides the next step. LangGraph implements this control flow (LangChain, Inc., 2024). To compose measurements, the base VLM either writes code for a deterministic Python executor or reasons over the returned values directly. In this way, it calculates distances, resolves turns, estimates object extents, and deduplicates instances across views. Figure 3 shows three such plans for questions about one video, and Appendix A.2 shows their full traces.

For VSIBench counting, SPLIT averages its tool count with the count the base VLM gives without tools.

The prompt defines coordinate and direction conventions. When measurements disagree, the base VLM can gather more evidence before it commits to an answer. Appendix Figure 7 shows a full trace in which two consistent measurements from different frames expose an incorrectly grounded object and change the final direction.

3.4 Ground-Truth Perception

To measure how much more accurate perception would help, we replace the perception tools' estimates on VSIBench, VSTIBench, and ReVSI with the ground-truth object regions, geometry, and camera information that each benchmark's annotations provide. MMSI-Bench and DSI-Bench provide no 3-D ground truth (scene point clouds or meshes, object boxes, or camera poses), so we run the ground-truth interventions only on these three benchmarks. The base VLM still chooses tools and composes their returned values into an answer. We evaluate only questions that SPLIT + GT can serve. Ground-truth tools report the annotated objects in the sampled frames, so SPLIT + GT measures the accuracy that perfect perception of those frames allows; on VSIBench counting, its only misses are objects that the frames never show and one annotation error. On VSIBench, SPLIT + GT also passes the base VLM the question's task and object labels, with a prompt that states each task's procedure (Appendix A.5).

4 Experimental Setup

Benchmarks and metrics. We evaluate on VSIBench (Yang et al., 2024) across all eight spatial tasks: object counting, absolute and relative distance, object and room size estimation, relative direction, route planning, and appearance order. We also evaluate on the official *Debiased* subset,

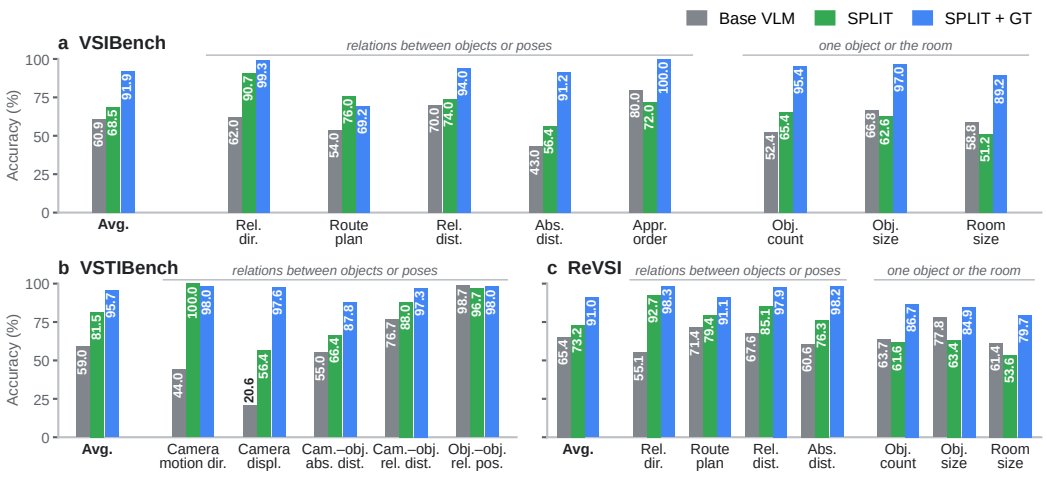


Figure 4: **Accuracy by Task.** Perception tools raise accuracy mainly on relations between objects or camera poses, while ground truth (SPLIT + GT) adds about twice as much to tasks that measure one object or the room as to relations. Each panel shows one benchmark's analysis subset (Section 4), and Avg. is each benchmark's score, the mean over its tasks.

which removes questions that admit non-visual shortcuts (Brown et al., 2025). ReVSI covers seven of these tasks, all except appearance order, and ties each question to an official frame sample in which a referenced object may be absent (Zhang et al., 2026). VSTIBench adds camera-motion questions, such as camera displacement and movement direction, to camera–object and object–object relations (Fan et al., 2025). To test transfer beyond static indoor video, we also evaluate on MMSI-Bench, which asks spatial questions about sets of unposed images (Yang et al., 2025b), and on DSI-Bench, which asks about videos of indoor and outdoor dynamic scenes (Zhang et al., 2025). Following each benchmark's protocol, we score multiple-choice questions by exact-match accuracy and numeric questions by mean relative accuracy (MRA), which gives partial credit according to relative error, and each benchmark's score averages its task scores.

Configurations. We use all models off the shelf, without fine-tuning. The base VLM is `gemini-3.1-pro` unless we name another model. We also run the open-weight base VLMs Gemma-4-31B and Qwen3.6-27B with the same sampling setup as SpatialClaw (Cho et al., 2026), and each serves as its own grounding VLM. We compare three conditions: the **base VLM** answers without perception tools, **SPLIT** uses perception tools, and **SPLIT + GT** uses perception tools that return ground-truth values. Due to computational constraints, we run the remaining analyses on *analysis subsets* of VSIBench, VSTIBench, and ReVSI. We selected design choices on VSIBench subsets before one full run of the seven non-counting tasks and a separate full run of object counting. Appendix A.5 gives the implementation details.

5 Results and Analysis

SPLIT improves the score on VSIBench, VSTIBench, and ReVSI, and each gain holds under paired resampling of scenes (Table 1). We examine which tasks benefit and how much accurate perception can recover. Together, these comparisons separate access to geometric evidence from the base VLM's ability to use it.

Finding 1: Geometry helps spatial relations more consistently than object extent

Relations and routes gain the most. Figure 4 groups the tasks by what their answers depend on: how several objects or camera poses are arranged, or the size or count of one object or the room. Relative direction and route planning improve the most on VSIBench and ReVSI (Appendix Tables 3

Table 1: **Accuracy on Five Spatial Benchmarks.** SPLIT raises accuracy over the base VLM on all five benchmarks, and ground-truth tools lift it to at least 90% on the three benchmarks with 3-D ground truth. Green parentheses give the gain over the base VLM, and bold and underlining mark the highest and second-highest scores in each column, excluding SPLIT + GT. *Evaluation setups differ: published VSTIBench rows use the original-release labels, while our rows use the corrected labels of the current release; our DSI-Bench rows use the 1,000-question subset of SpatialClaw (Cho et al., 2026); and SpatialClaw uses Qwen3.6-27B as its base VLM and reports 1,000 questions each from the *Debiased* subset of VSIBench and from VSTIBench. †SPLIT + GT, whose tools return ground-truth measurements, runs on each benchmark's analysis subset (Table 2); MMSI-Bench and DSI-Bench provide no 3-D ground truth. Appendix Tables 3, 4, and 5 give per-task scores.

Method	VSIBench	VSTIBench*	ReVSI	MMSI-Bench	DSI-Bench*
Proprietary					
GPT-4o	34.0	38.2	—	30.3	37.2
Gemini-2.5 Pro	51.5	42.4	—	36.9	46.9
GPT-5	—	—	—	41.9	40.1
GPT-5.2	49.2	—	50.9	—	—
Gemini 3 Pro	60.5	—	60.9	—	—
Open-weight					
InternVL3.5-38B	60.8	—	54.1	—	39.1
LLaVA-Video-72B	39.7	44.0	37.8	—	—
Supervised spatial specialists					
VLM-3R-7B (Fan et al., 2025)	60.9	58.8	50.1	—	—
Cambrian-S-7B (Yang et al., 2025a)	67.5	—	49.1	—	—
Cambrian-P (Yang et al., 2026)	**73.7**	68.9	—	—	—
Training-free					
TRACE (Hua et al., 2026)	57.8	—	—	—	—
SpatialClaw* (Cho et al., 2026)	57.9	<u>70.3</u>	—	<u>53.9</u>	**68.1**
Base VLM (Gemini 3.1 Pro)	60.4	57.7	<u>63.8</u>	49.3	50.2
SPLIT (Ours)	<u>67.9</u> (+7.5)	**78.3** (+20.6)	**71.8** (+8.1)	**60.9** (+11.6)	<u>65.0</u> (+14.8)
SPLIT + GT†	91.9	95.7	91.0	—	—

Table 2: **Matched Comparison Across Base VLMs.** SPLIT raises accuracy in most matched comparisons, and ground-truth tools raise every base VLM further. The VSIBench, VSTIBench, and ReVSI columns use the analysis subsets (Section 4). Parentheses give the change from the base VLM (the same model without tools), green for a gain and red for a loss. Dashes mark results that are not available. †SPLIT + GT covers only the questions whose ground truth it can serve.

Model	VSIBench *500 q.*	VSTIBench *450 q.*	ReVSI *999 q.*	MMSI-Bench *1,000 q.*	DSI-Bench *1,000 q.*
Gemini 3.1 Pro					
Base VLM	60.9	59.0	65.4	49.3	50.2
SPLIT	68.5 (+7.7)	81.5 (+22.5)	73.2 (+7.8)	60.9 (+11.6)	65.0 (+14.8)
SPLIT + GT	91.9 †	95.7	91.0 †	—	—
Gemma-4-31B					
Base VLM	53.7	52.4	55.8	37.4	45.0
SPLIT	65.8 (+12.1)	77.1 (+24.7)	58.6 (+2.8)	50.5 (+13.1)	56.8 (+11.8)
SPLIT + GT	86.8 †	89.6	94.8 †	—	—
Qwen3.6-27B					
Base VLM	59.2	58.9	56.2	40.9	48.8
SPLIT	65.4 (+6.1)	69.5 (+10.6)	47.1 (−9.0)	43.6 (+2.7)	57.4 (+8.6)
SPLIT + GT	84.7 †	88.8 †	92.6 †	—	—

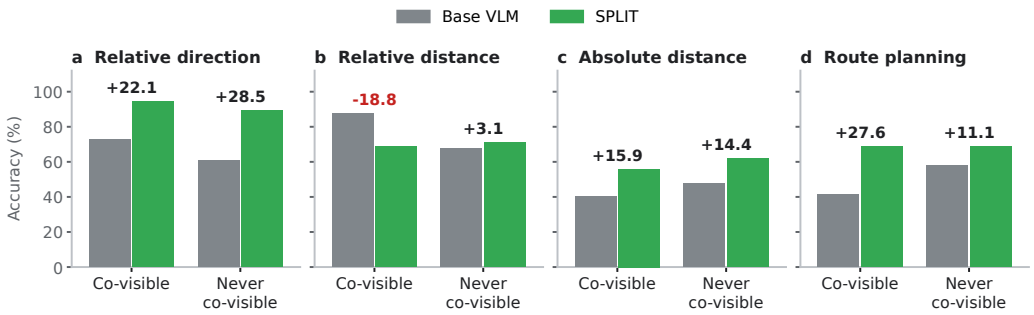


Figure 5: **Gains by Co-visibility.** The tools help more on relation questions whose objects never appear together in a sampled frame. We split each multi-object VSIBench task by whether any of the sampled frames shows every object the question names, using ground-truth geometry to decide visibility. Numbers above the bars give SPLIT's gain over the base VLM.

and 5). These tasks ask the base VLM to relate locations in one world coordinate system, even for objects that never appear in the same frame (Figure 1). Across VSIBench relation questions, SPLIT gains 22.1% over the base VLM when the objects a question names never share a sampled frame, against 11.9% when one frame shows them all (Appendix Table 8). The gain grows because the base VLM loses 13.8% when the objects never share a frame, while SPLIT loses only 3.6%. Figure 5 breaks this down by task: relative direction and relative distance both show this pattern. The tools help about equally on absolute distance, where the base VLM estimates metric distance poorly even when both objects are in view. The harness therefore helps most on relations that no single frame shows. Relative direction, for example, depends on the arrangement of objects but is unchanged by a uniform rescaling of the scene. Object-size questions instead require accurate physical extents, and the tools reduce their scores on VSIBench and ReVSI (Figure 4). This contrast suggests that a reconstruction can support useful spatial relations while still measuring object boundaries poorly. The tools also lower ReVSI room size and counting. The VSIBench gain also persists on the *Debiased* subset, so it does not depend solely on questions that admit non-visual shortcuts.

Camera motion benefits from an explicit trajectory. On VSTIBench, camera-motion direction and camera displacement gain the most (Figure 4b; Appendix Table 4). The tools expose camera poses in the world coordinate system, giving the base VLM direct inputs for these calculations. Yet object-object relative position regresses slightly even as the score rises. Explicit geometry therefore helps particular operations; it does not improve every relation that can be expressed in 3-D.

The harness transfers to images, motion, and other base VLMs. On MMSI-Bench, the harness also improves accuracy on sets of unposed images (Table 1). On DSI-Bench, most of the gain comes from temporally reversed clips, while clips with moving objects and a static camera regress. This pattern is consistent with how the harness uses camera geometry and reconstructs the scene as rigid, but the comparison does not isolate that mechanism. With open-weight base VLMs, SPLIT raises accuracy in nine of ten matched comparisons, by 2.7% to 24.7%, and lowers only Qwen3.6-27B's ReVSI score, by 9.0% (Table 2).

Finding 2: Accurate perception recovers most of the remaining gap

Ground truth raises accuracy. SPLIT + GT replaces the values returned through the tool interface, so it covers only the questions whose ground truth it can serve. On VSTIBench, where the prompt also stays fixed, ground truth raises SPLIT's accuracy from 81.5% to 95.7%, most on camera-object absolute distance and camera displacement, so this headroom lies in the supplied evidence (Figure 4b). On VSIBench, ground truth raises SPLIT's accuracy from 68.5% to 91.9% under a protocol that also supplies the task, target labels, and per-task procedures (Section 3.4). VSIBench route planning is the exception, reaching only 69.2% under ground truth. Ground-truth tools also raise both open-weight base VLMs: across the three benchmarks, Gemma-4-31B scores 86.8% to 94.8% and Qwen3.6-27B 84.7% to 92.6% (Table 2).

Ground truth helps most where the tools help least. The tools raise accuracy on relations between objects or camera poses by an average of 12.0% on VSIBench and 19.7% on ReVSI, but they barely change accuracy on the tasks that measure one object or the room on VSIBench and lower it on ReVSI (Figure 4). On these tasks, ground truth raises SPLIT's accuracy by an average of 34.1% on VSIBench and 24.2% on ReVSI, about twice its gain on relations. Tools that measure extents and count instances more accurately would therefore help these tasks most.

6 LIMITATIONS

SPLIT requires substantial inference-time computation: on the VSIBench analysis subset, it calls the base VLM a median of 16 times and generates a median of $5{,}980$ tokens, including thinking tokens, per question. VSIBench, VSTIBench, and ReVSI derive from static indoor scans; DSI-Bench covers a narrower set of motion questions, so we have not tested the harness on unconstrained video of dynamic scenes. The ground-truth interventions replace several perception components together and cover only questions whose ground truth they can serve, so they measure the combined headroom of perception. They do not establish that perception alone limits SPLIT.

7 CONCLUSION

To answer spatial questions about a video, a model must recover scene geometry from pixels and reason with that geometry. SPLIT separates these stages through a shared tool interface: perception models supply measurements, and a base VLM composes them to answer questions. The harness improves accuracy on all five benchmarks, with more consistent gains for spatial relations and camera motion than for object extents. Ground-truth interventions on VSIBench, VSTIBench, and ReVSI show that accurate perception closes most of the gap to a perfect score.

AI USE STATEMENT

We designed the experiments and method. We used generative AI to write the experimental code, launch experiments on GPUs, and write the visualization code behind the figures. We also used it to edit the manuscript. We carefully checked all scripts and verified every quantitative claim against the raw experiment outputs. We take responsibility for the content of this work.

REFERENCES

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. *arXiv preprint arXiv:2511.04655*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung Chen. SpatialClaw: Rethinking action interface for agentic spatial reasoning. *arXiv preprint arXiv:2606.13673*, 2026.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Tanmay Gupta and Aniruddha Kembhavi. 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.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. *arXiv preprint arXiv:2603.23404*, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

LangChain, Inc. LangGraph. `https://github.com/langchain-ai/langgraph`, 2024.

Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify Anything: Scaling indoor 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22225–22233, 2025.

Bharath Raj Nagoor Kani and Noah Snavely. G3T up! gravity aligned coordinate frames simplify pointmap processing. *arXiv preprint arXiv:2605.27372*, 2026.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

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. *arXiv preprint arXiv:2412.14171*, 2024.

Jihan Yang, Zifan Zhao, Xichen Pan, Shusheng Yang, Junyi Zhang, Bingyi Kang, Hu Xu, Shang-Wen Li, and Saining Xie. Cambrian-P: Pose-grounded video understanding. *arXiv preprint arXiv:2605.22819*, 2026.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025a.

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. *arXiv preprint arXiv:2505.23764*, 2025b. URL `https://arxiv.org/abs/2505.23764v2`.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. ReVSI: Rebuilding visual spatial intelligence evaluation for accurate assessment of VLM 3D reasoning. *arXiv preprint arXiv:2604.24300*, 2026.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. DSI-Bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025. URL https://arxiv.org/abs/2510.18873v1.

A Appendix

A.1 Per-task results

Tables 3, 4, 5, 6, and 7 give the per-task scores behind Table 1; the first three also list published results under their own protocols.

Table 3: **VSIBench Accuracy by Task.** SPLIT gains most over the base VLM in relative direction, route planning, and absolute distance. The Train column marks training-free methods (TF) and supervised fine-tuning (SFT). Each Δ row is SPLIT minus the base VLM on the same questions; the full-benchmark object-size Δ has a 95% scene-clustered interval from −9.5% to −4.9%. Bold and underlining mark the best and second-best full-benchmark scores. Human and general-purpose results come from VSI-Bench and later comparison tables (Yang et al., 2024; 2025a; Zhang et al., 2026). ReVSI (Zhang et al., 2026) reports no appearance-order score for GPT-5.2, Gemini 3 Pro, InternVL3.5-38B, or LLaVA-Video-72B, so their averages are the published eight-task scores. ‡SPLIT + GT covers the analysis-subset questions that its run can serve. Eleven route-planning questions from ARKitScenes have no usable annotations and count as wrong.

			Numerical				Multiple choice			
Method	Train	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Human (Yang et al., 2024)	—	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-2.5 Pro	TF	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
GPT-5.2	TF	49.2	57.1	33.4	64.6	59.0	48.0	33.3	36.7	—
Gemini 3 Pro	TF	60.5	45.3	38.3	73.0	47.4	70.0	60.8	65.3	—
InternVL3.5-38B	TF	60.8	73.9	39.2	73.0	65.0	66.2	72.0	36.1	—
LLaVA-Video-72B	TF	39.7	51.9	24.0	57.4	32.7	42.4	37.4	32.0	—
Supervised spatial specialists										
VLM-3R-7B (Fan et al., 2025)	SFT	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
Cambrian-S-7B (Yang et al., 2025a)	SFT	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
Cambrian-P (Qwen2.5-7B) (Yang et al., 2026)	SFT	**73.7**	**74.9**	**60.1**	**76.0**	**76.9**	**74.8**	89.5	52.6	**85.0**
Training-free methods										
TRACE (Gemini 3 Pro) (Hua et al., 2026)	TF	57.8	47.6	38.8	73.9	45.6	63.9	61.7	58.0	73.0
Full benchmark										
Base VLM	TF	60.4	51.2	43.6	73.3	53.1	69.2	62.4	58.8	71.7
SPLIT (Ours)	TF	67.9	58.3	58.3	66.1	56.3	69.9	**90.0**	**73.7**	70.7
Δ		+7.5	+7.1	+14.7	-7.2	+3.2	+0.7	+27.6	+14.9	-1.0
Debiased subset										
Base VLM	TF	57.5	46.4	41.2	62.9	54.7	69.0	62.3	51.8	71.7
SPLIT (Ours)	TF	66.8	58.8	56.4	60.2	57.5	71.6	86.7	71.1	72.3
Δ		+9.3	+12.5	+15.2	-2.7	+2.8	+2.6	+24.4	+19.3	+0.6
Analysis subset										
Base VLM	TF	60.9	52.4	43.0	66.8	58.8	70.0	62.0	54.0	80.0
SPLIT (Ours)	TF	68.5	65.4	56.4	62.6	51.2	74.0	90.7	76.0	72.0
Δ		+7.7	+13.0	+13.4	-4.2	-7.6	+4.0	+28.7	+22.0	-8.0
SPLIT (Ours) + GT‡	TF	91.9	95.4	91.2	97.0	89.2	94.0	99.3	69.2	100.0

A.2 Full traces

Each example question in the paper's figures has its full recorded episode here, with the base VLM's plan and code on the left and the tools' outputs on the right. Gray marks the plan, green marks steps with correct results, and red marks steps with wrong results.

A.3 Diagnostic question sets

We interpret each diagnostic result in the main text on a fixed population stated in advance. *Full benchmark* (5,130 questions): the complete VSIBench evaluation population; the base-VLM reference covers all 5,130 questions. *Debiased subset* (2,362 questions): the published shortcut-pruned subset

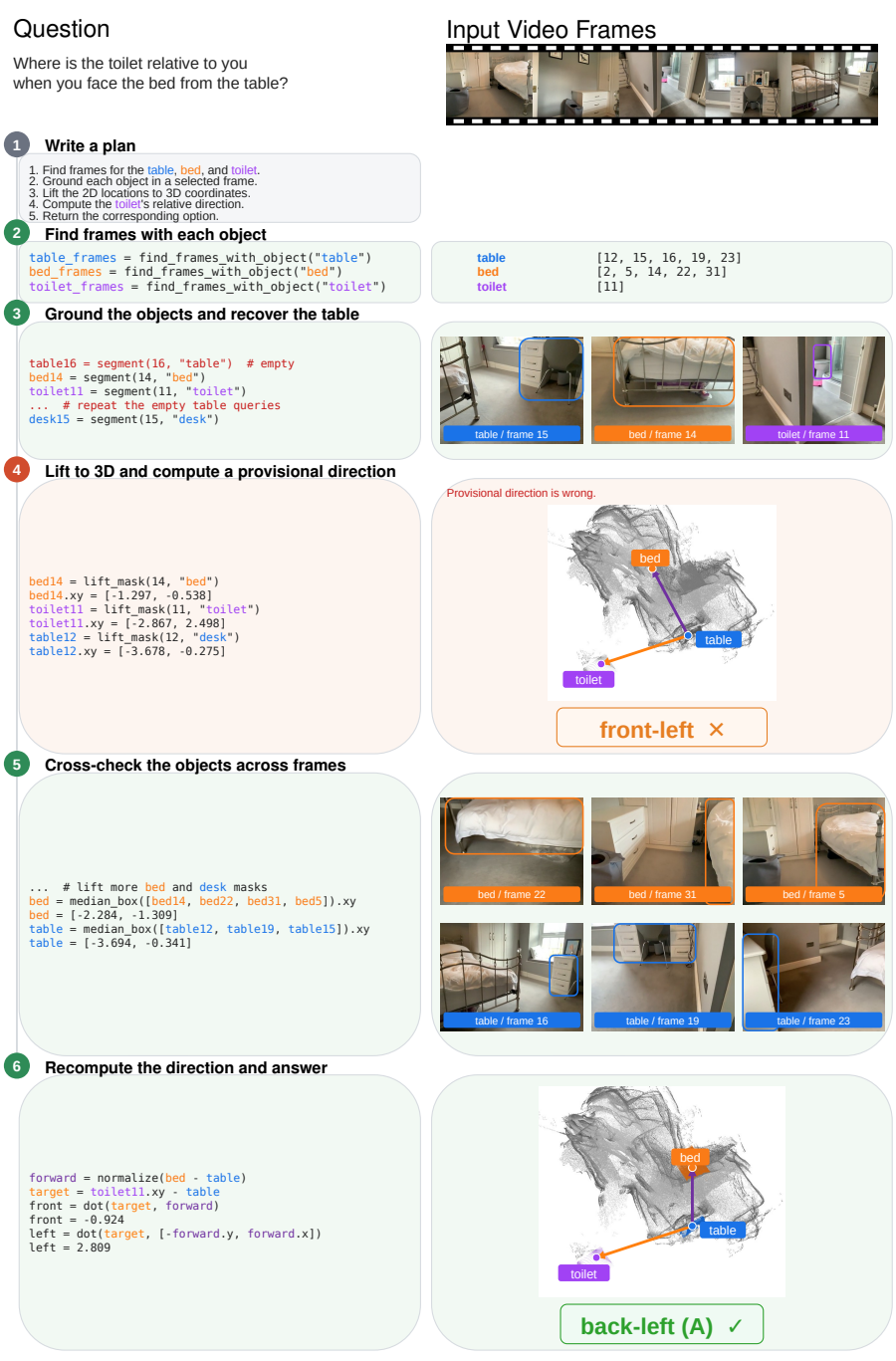


Figure 6: **A Full SPLIT Episode: The Figure 1 Question.** Measurements from more frames correct a provisional *front-left* result to the correct answer, *back-left*. Outlines mark recorded detection boxes, and the top-down views place the recorded centroids on the episode's reconstruction.

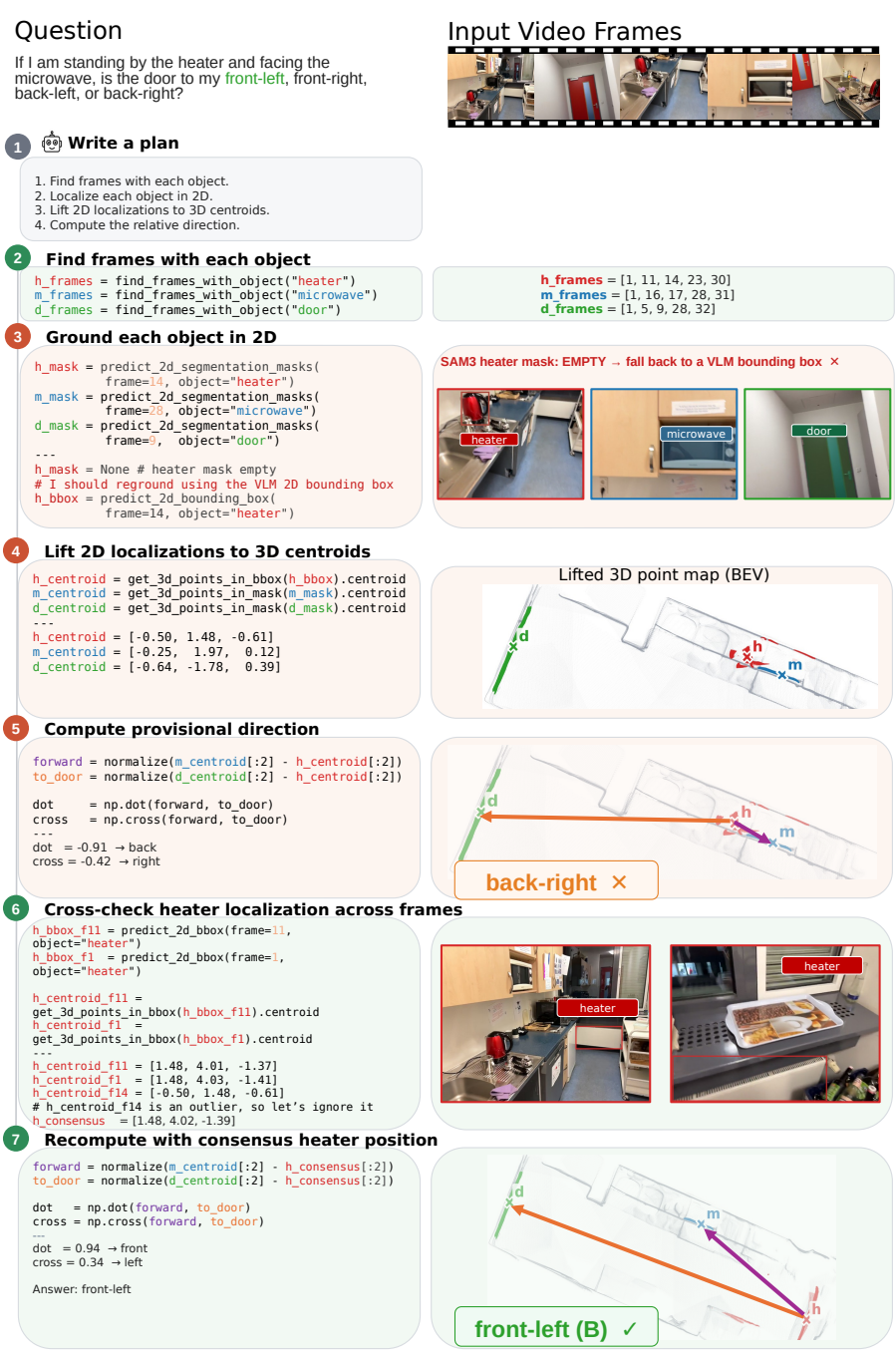


Figure 7: **A Full SPLIT Episode.** Cross-checking measurements from several frames lets the base VLM catch a wrongly grounded object before it answers. Steps 1–7 show the recorded episode for the question of Figure 2, in which the fallback grounding VLM boxes a kettle instead of the heater in Step 3.



Figure 8: **A Full SPLIT Episode: Relative Direction.** Centroids lifted from masks and a cross-check with 2-D points give the same correct answer, *front-right*. Steps 1–7 show the recorded episode behind Figure 3(a). Rectangles mark the recorded segmentation boxes, and the top-down views plot recorded centroids over the episode's reconstruction.

Table 4: **VSTIBench Accuracy by Task.** SPLIT raises VSTIBench accuracy over the base VLM, mostly through camera motion direction and camera displacement. The Δ row is SPLIT minus the base VLM on all questions, and MRA marks tasks scored by mean relative accuracy. The GPT-4o, Gemini-2.5 Pro, and LLaVA-Video-72B rows come from the VSTIBench release (Fan et al., 2025). Bold and underlining mark the best and second-best full-benchmark scores. *Published rows use the original-release labels, while our rows use the corrected camera-displacement and camera–object-distance labels of the current release. †These rows compare the three conditions on the analysis subset (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**							
GPT-4o	TF	38.2	29.5	23.4	37.3	58.1	42.5
Gemini-2.5 Pro	TF	42.4	29.2	8.0	30.0	84.0	60.7
LLaVA-Video-72B	TF	44.0	32.3	10.5	48.1	78.3	50.9
VLM-3R-7B (Fan et al., 2025)	SFT	58.8	39.4	39.6	60.6	86.5	68.6
Cambrian-P (Yang et al., 2026)	SFT	<u>68.9</u>	42.5	<u>46.6</u>	<u>87.7</u>	**94.3**	73.2
Full benchmark							
Base VLM	TF	57.7	<u>49.0</u>	18.5	45.3	<u>93.5</u>	<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
Analysis subset†							
Base VLM	TF	59.0	55.0	20.6	44.0	98.7	76.7
SPLIT (Ours)	TF	81.5	66.4	56.4	100.0	96.7	88.0
SPLIT (Ours) + GT	TF	95.7	87.8	97.6	98.0	98.0	97.3

used for the second dataset view in Table 3; the base-VLM cell is the same base-VLM run rescored on all 2,362 subset questions, which scores 57.5%.

A.4 Extended results: corroboration, robustness, and implementation

We print scores and differences to one decimal place, rounding half up from unrounded values, and we compute each difference before rounding.

Perception constants. The deployed configuration inherits upstream SAM3 video default filter semantics; applies Tukey's $1.5 \times$ IQR rule to axis-aligned-bounding-box endpoints before extent aggregation; and groups lifted 3-D observations with scikit-learn-default Euclidean DBSCAN (`min_samples`=5, native noise labels) at `eps`=0.5 m, a radius chosen because indoor objects and furniture typically occupy meter and sub-meter spatial scales. No constant inspects or depends on an answer; a few generic checks that reject implausible outputs, such as points below the floor, are enabled per task.

Grounding accuracy. Table 10 compares the objects that the perception tools find with frame-level object annotations: the frame scan finds almost every requested object in at least one frame (recall 0.92), but per-frame grounding and tool-based multi-view counting recall range from 0.68 to 0.81.

Tasks where the base VLM is stronger. Tools do not win every task; the losses bound where explicit geometry helps. On the *Debiased* subset (Table 3), SPLIT sits *below* its base VLM only on object size. It leads on relative distance, appearance order, room size, and object counting. Its counting score is 58.8%, compared with 46.4% for the base VLM.

Second-benchmark check (VSTIBench). To test whether the perception bottleneck is specific to VSIBench, we also evaluate SPLIT on VSTIBench (Fan et al., 2025), a spatial-*temporal* video benchmark on the same ScanNet, ScanNet++, and ARKitScenes sources (Table 4). We evaluate all 5,736 questions of the benchmark's current official release. On the full benchmark SPLIT scores

Table 5: **ReVSI Accuracy by Task.** SPLIT raises ReVSI accuracy over the base VLM, gaining most in relative direction and route planning and losing in counting, object size, and room size. The Δ row compares SPLIT with the base VLM; published scores come from ReVSI (Zhang et al., 2026), and the Frames column lists each row's frame count or frame rate. Bold and underlining mark the highest and second-highest listed full-benchmark scores. †The analysis subset is the official small subset of ReVSI with its benchmark-provided 32 frames (Section 4); SPLIT + GT covers its annotated questions.

				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	73.9	**63.0**	48.4	34.9	38.2
Gemini 3 Pro	TF	1 FPS	60.9	<u>60.1</u>	54.7	**79.3**	51.9	<u>68.1</u>	56.0	56.4
InternVL3.5-38B	TF	64	54.1	43.8	60.6	70.2	58.4	57.4	45.9	42.7
LLaVA-Video-72B	TF	64	37.8	40.1	29.6	59.3	27.9	39.6	24.8	43.0
VLM-3R-7B (Fan et al., 2025)	SFT	32	50.1	41.6	<u>61.6</u>	64.8	52.5	46.5	49.5	34.1
Cambrian-S-7B (Yang et al., 2025a)	SFT	128	49.1	48.4	60.5	65.5	46.7	37.1	48.5	37.0
Full benchmark										
Base VLM	TF	32	<u>63.8</u>	**64.6**	59.6	<u>76.8</u>	<u>61.9</u>	67.0	<u>58.0</u>	<u>58.5</u>
SPLIT (Ours)	TF	32	**71.8**	59.9	**66.9**	65.5	49.8	**84.1**	**93.2**	**83.3**
Δ			+8.1	-4.7	+7.2	-11.3	-12.0	+17.1	+35.3	+24.8
Analysis subset†										
Base VLM	TF	32	65.4	63.7	60.6	77.8	61.4	67.6	55.1	71.4
SPLIT (Ours)	TF	32	73.2	61.6	76.3	63.4	53.6	85.1	92.7	79.4
SPLIT (Ours) + GT	TF	32	91.0	86.7	98.2	84.9	79.7	97.9	98.3	91.1

Table 6: **MMSI-Bench Accuracy by Category.** SPLIT raises accuracy in every MMSI-Bench category, most on attributes and motion. Both rows use Gemini 3.1 Pro on all MMSI-Bench questions; Overall is the accuracy over all questions, and the Δ row compares SPLIT with the base VLM.

Method	**Overall**	Positional Relationship	Attribute	Motion	Multi-Step Reasoning
Base VLM	49.3	50.8	48.5	46.0	48.5
SPLIT (Ours)	60.9	61.3	66.2	62.7	55.1
Δ	+11.6	+10.5	+17.7	+16.7	+6.6

Table 7: **DSI-Bench Accuracy by Category.** SPLIT gains most on camera motion, object–camera orientation, and temporally reversed clips, and loses on object motion seen by a static camera. Both rows use Gemini 3.1 Pro on the DSI-Bench questions of Table 1; Overall is the accuracy over all questions, the right columns group the same questions by how the clip is played, and the Δ row compares SPLIT with the base VLM.

		Object motion		Camera motion		Object–camera		Clip playback			
Method	**Overall**	Static camera	Moving camera	Static scene	Dynamic scene	Distance	Orientation	Standard	Reversed	Flipped	Reversed and flipped
Base VLM	50.2	53.9	49.6	52.1	43.1	76.4	47.7	66.5	37.1	63.3	33.3
SPLIT (Ours)	65.0	49.4	53.9	87.3	66.1	81.9	77.3	68.1	63.7	63.3	64.7
Δ	+14.8	-4.5	+4.3	+35.2	+23.0	+5.6	+29.5	+1.5	+26.5	0.0	+31.3

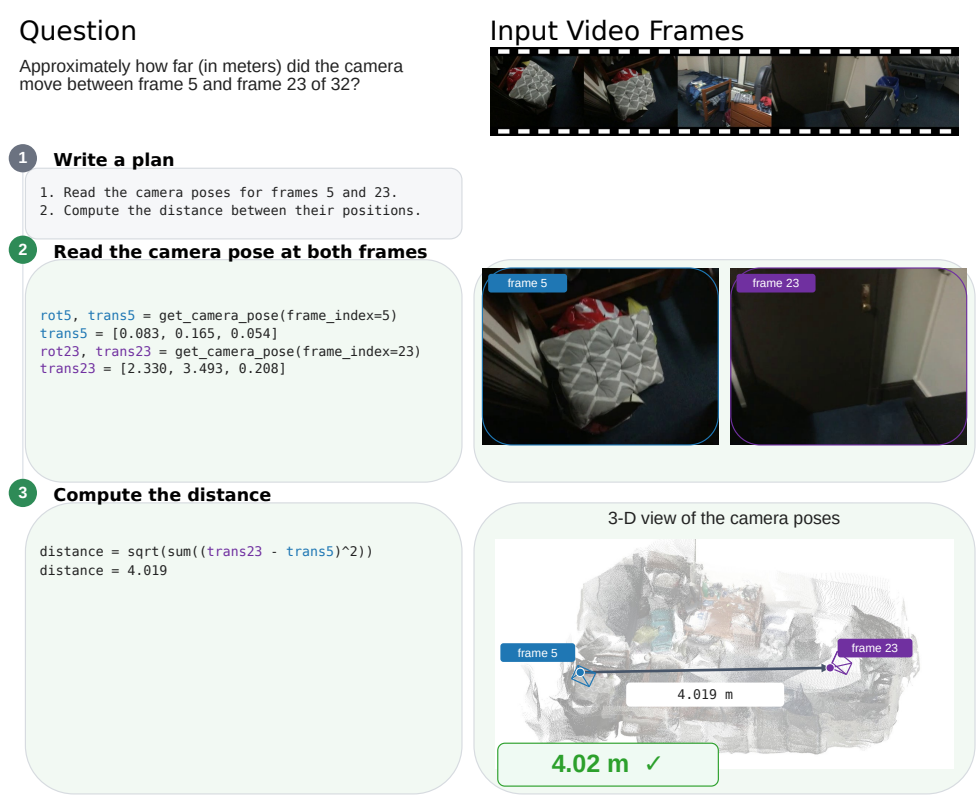


Figure 9: **A Full SPLIT Episode: Camera Displacement.** The base VLM answers from the distance between two returned camera translations. Steps 1–3 show the recorded episode behind Figure 3(b). The same reconstruction supplies the camera poses.

Table 8: **Gains When the Named Objects Never Share a Frame.** SPLIT's improvement over the base VLM nearly doubles on relation questions whose named objects never appear in one frame. *Co-visible* means that at least one of the 32 frames shows all the named objects, and the table omits the 119 questions whose visibility is unresolved. Scores give each relation question type equal weight.

Named objects	Questions	Base VLM	SPLIT	Gain
Co-visible	237	76.3	88.2	+11.9
Never co-visible	1,322	62.5	84.6	+22.1
Difference				+10.2

78.3%, versus 57.7% for the base VLM on the same questions. On the VSTIBench analysis subset, SPLIT + GT raises accuracy from 81.5% to 95.7%, improving four of the five tasks (Table 4).

Third-benchmark check (ReVSI). On all 6,158 ReVSI questions at 32 frames, SPLIT scores 71.8% against 63.8% for the base VLM, a +8.1% gain (per-task scores in Table 5).

Relation questions whose objects never share a frame. To test whether the harness matters most when no single frame shows the answer, we split each multi-object VSIBench task by whether any of the 32 frames shows all the objects a question names (Table 8 and Figure 5). Ground-truth instance points, camera poses, and depth decide whether each object is visible in each frame. We leave out appearance order, which asks when each object first appears. On relation questions, SPLIT's gain over the base VLM is 10.2 points larger when the objects never appear together than when one frame shows them all. Figure 14 repeats the comparison for each base VLM on the analysis subset, which also has a ground-truth perception run.

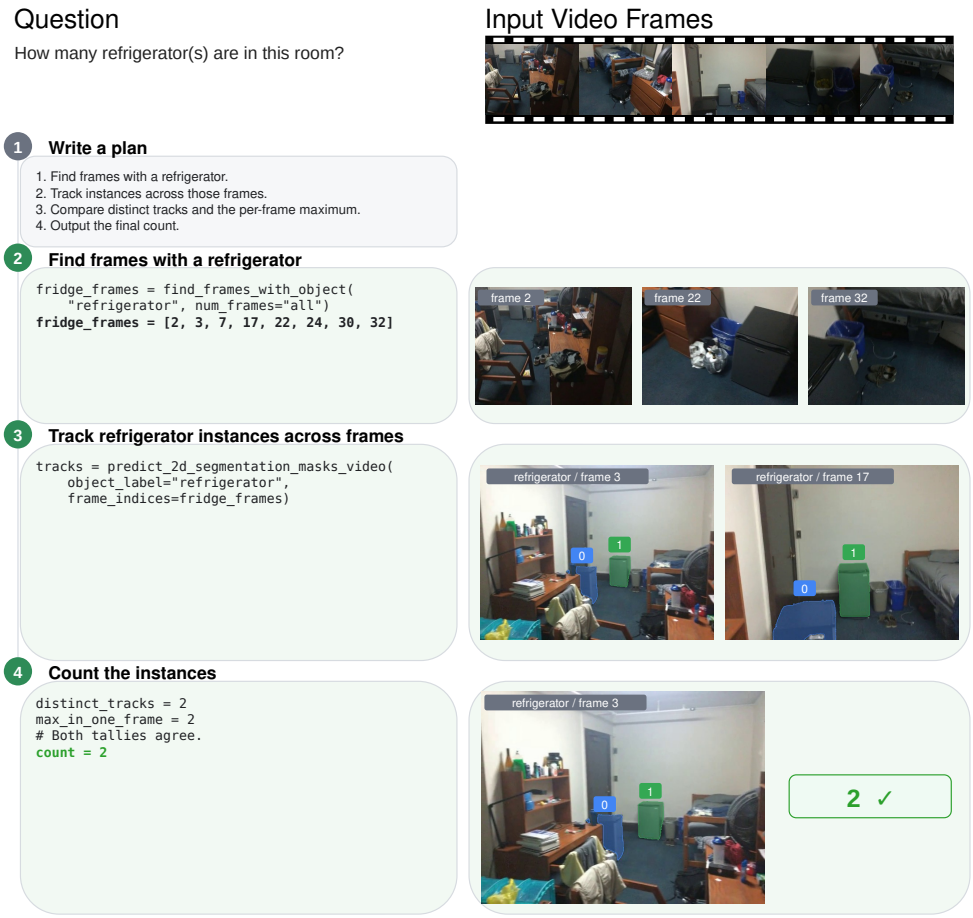


Figure 10: **A Full SPLIT Episode: Object Counting.** Tracking the refrigerators across frames gives a count of two. Steps 1–4 show the recorded episode behind Figure 3(c).

Table 9: **SPLIT Tools.** Every perception tool returns a measurement that does not depend on the task. The table omits one tool that renders the reconstruction from a new viewpoint, which the released harness drops.

Role	Tools
Frame retrieval	`find_frames_with_object`, `get_frame_image`
2-D grounding	`predict_2d_bounding_box`, `predict_2d_points`, `predict_2d_segmentation_masks`, `..._video`
2-D→3-D lift	`get_world_3d_point_from_2d`, `get_3d_points_in_bbox`, `get_3d_points_in_mask`
3-D geometry	`aggregate_aabbs_median/_robust`, `apply_coordinate_transform`, `get_camera_pose`, `query_3d_bbox_catalogue`, `cluster_3d_points`
Rendering	`render_topdown_bev`
Python execution	`execute_python_code`

A.5 Implementation Details

Grounding. SAM3 segments a named object in one frame or tracks its instances across frames. For the Gemini base VLMs, the `gemini-3.5-flash` grounding VLM, run at temperature 0, retrieves the frames likely to show an object, draws boxes, and places points. No tool switches between the two on its own: the prompt asks the base VLM to prefer masks for measurements and, when SAM3 finds no instance, to retry with a synonym or request a box.

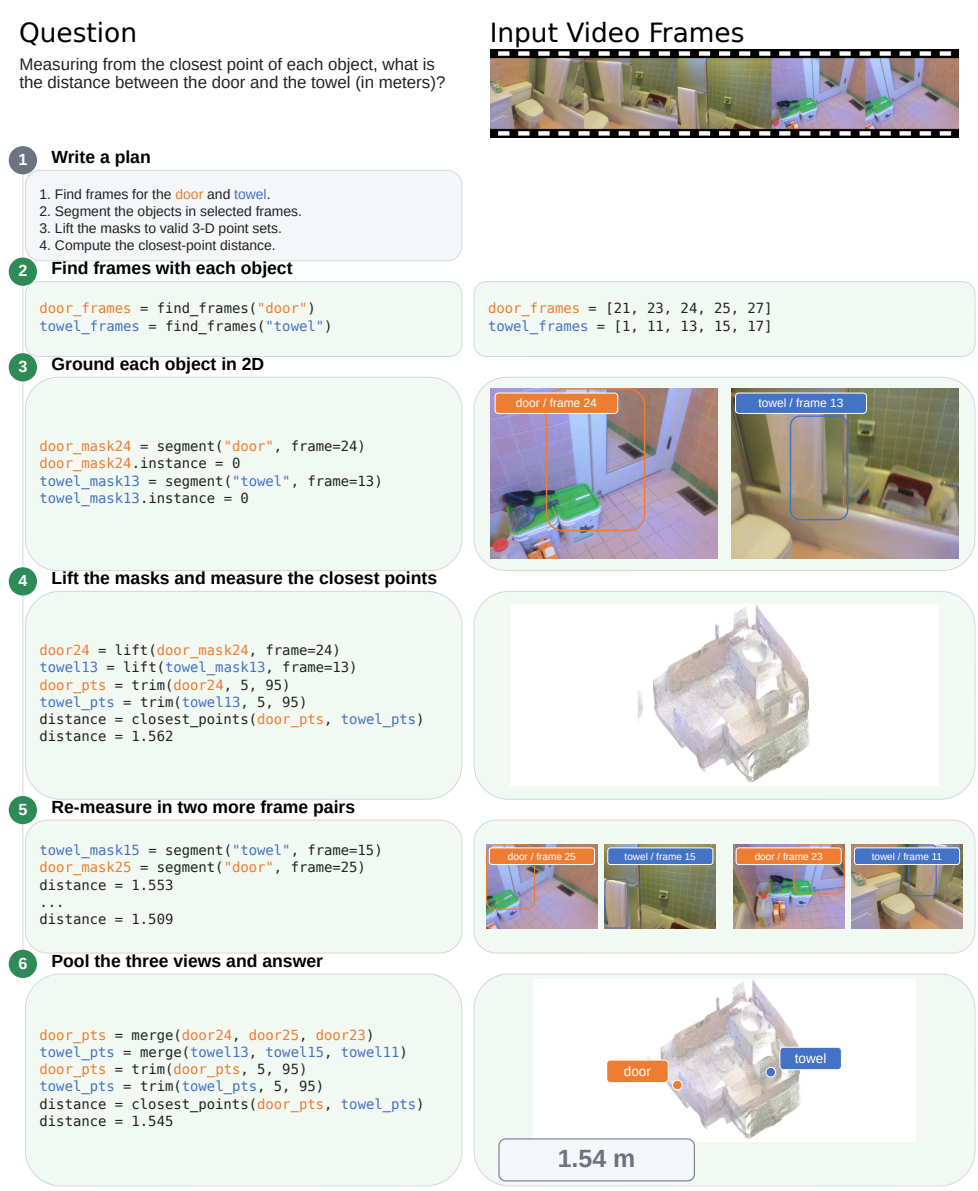


Figure 11: **A Full SPLIT Episode: Absolute Distance.** The base VLM measures the gap in three frame pairs and pools the objects' points into one answer. Steps 1–6 show the recorded episode, whose ground truth is 1.3 m. The outlines show recorded segmentation boxes, and the 3-D views show the episode's reconstruction.

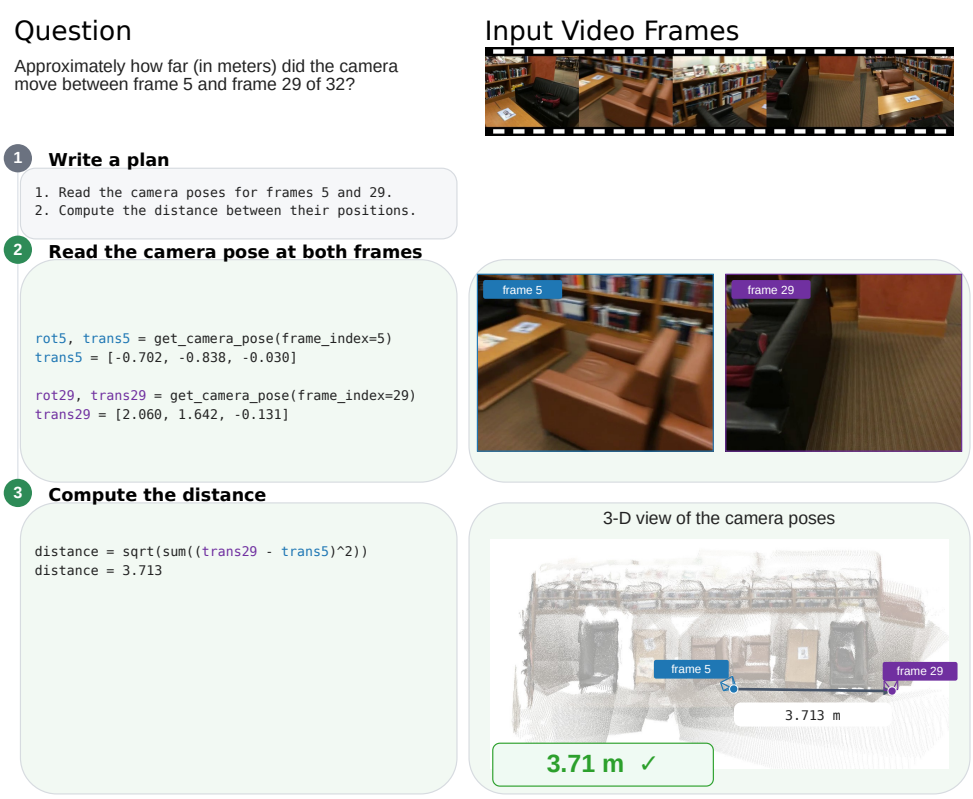


Figure 12: **A Full SPLIT Episode: Camera Displacement Between Frames 5 and 29.** The base VLM answers from the distance between two returned camera translations. Steps 1–3 show the recorded episode. The same reconstruction supplies the camera poses.

Reconstruction, scale, and caching. G3T predicts camera poses and per-frame 3-D points for the 32 frames in gravity-aligned coordinates, which we store with Z up and the floor in the XY plane. UniK3D predicts metric depth for the same frames. For each frame, we divide UniK3D's median depth by G3T's median depth, each taken over that model's valid pixels, and we use the median of these ratios as the scene's scale. The scale multiplies G3T's points, depths, and camera translations and leaves its rotations unchanged. In every reported run, all tools read this reconstruction, even when the base VLM asks a tool for another source. The `query_3d_bbox_catalogue` tool reads class-agnostic 3-D boxes that the RGB-only Cubify Transformer (CuTR) (Lazarow et al., 2025) predicts once per scene from the 32 video frames and the camera intrinsics of a MapAnything reconstruction of the same video (Keetha et al., 2025), and it places those boxes in the scene with G3T's camera poses. We compute the reconstruction once per scene before evaluation. Within a run, repeated SAM3 requests for the same frame and label, and repeated frame-retrieval requests for the same label, return cached results.

A.6 Other base VLMs

Open-weight base VLMs. Gemma-4-31B and Qwen3.6-27B use the same perception tools, prompt, and 32 sampled frames as the Gemini base VLM, but they serve as their own grounding VLMs. Each decodes with thinking enabled and samples at temperature 0.6 with top-p 0.95 and top-k 64 for Gemma-4-31B and 20 for Qwen3.6-27B, following SpatialClaw (Cho et al., 2026). Table 2 reports their scores.

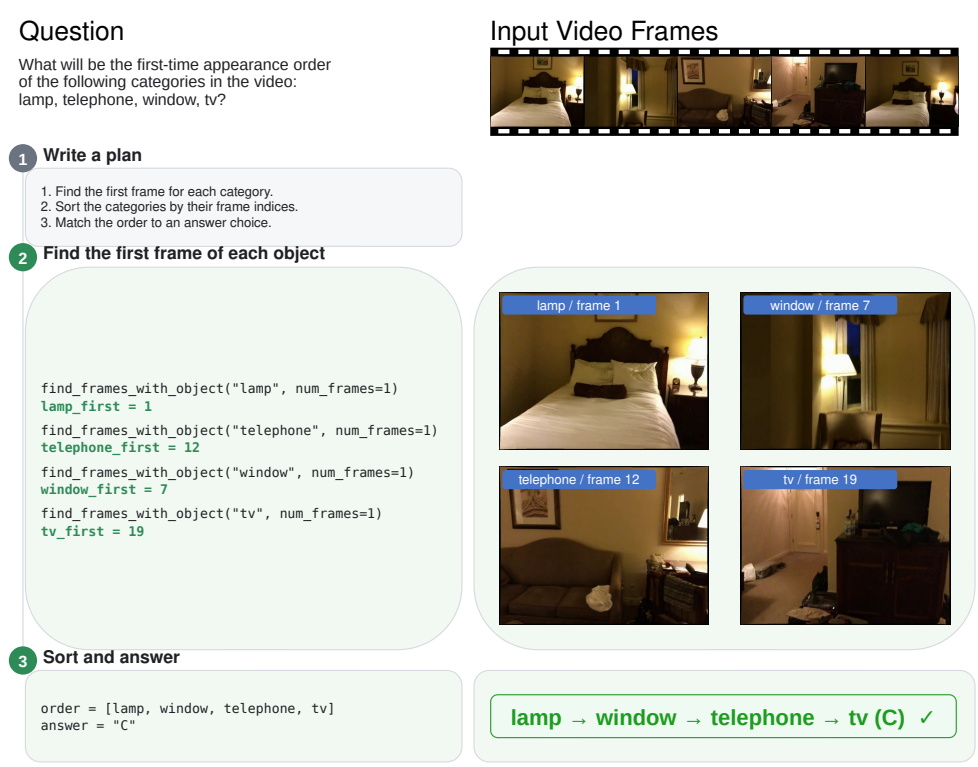


Figure 13: **A Full SPLIT Episode: Appearance Order.** Four first-appearance lookups give the correct order. Steps 1–3 show the recorded episode.

Table 10: **Grounding-Tool Accuracy.** The perception tools usually find at least one frame containing a requested object, but they miss or duplicate objects often enough to limit counting. Precision (P), recall (R), and F1 compare the detections of SPLIT's tool calls with frame-level object annotations on 172 ScanNet and 225 ARKitScenes questions. Found-any recall is the fraction of requested objects that the frame scan finds in at least one frame. The table rounds precision and recall for display and computes F1 from the unrounded counts.

Tool	Role	P	R	F1
`find_frames_with_object`	retrieval (found-any)	—	0.92	—
`predict_2d_points`	grounding	0.72	0.81	0.76
`predict_2d_bounding_box`	grounding	0.60	0.78	0.68
`predict_2d_segmentation_masks`	grounding	0.67	0.69	0.68
SAM3-video (max-frame)	counting dedup	0.80	0.68	0.73
SAM3-video (union)	counting dedup	0.67	0.78	0.72

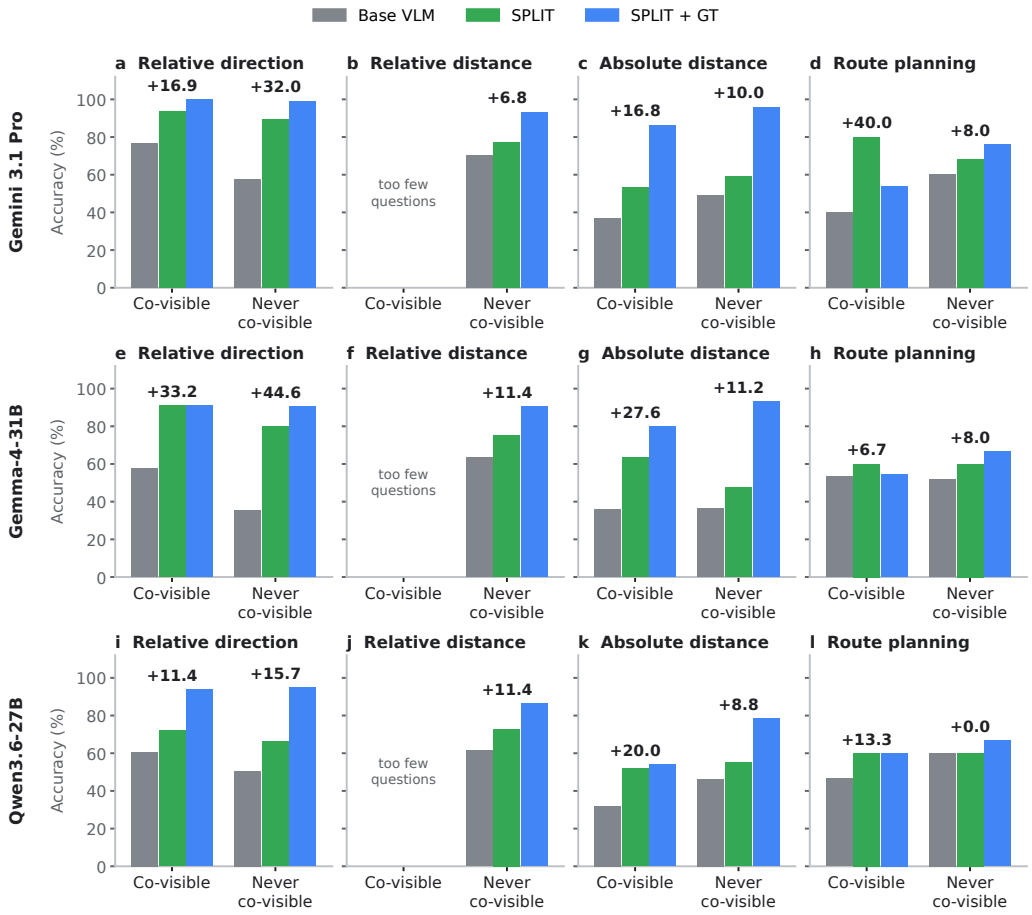


Figure 14: **Co-visibility Gains per Base VLM.** Every base VLM gains on relative direction whether or not the objects share a frame, and with ground-truth perception every base VLM answers those questions with at least 90% accuracy either way. Each row repeats Figure 5 on the analysis subset for one base VLM, and SPLIT + GT replaces estimated perception with ground truth. Too few relative-distance questions in this subset are co-visible (fewer than 5), so we do not report those bars, and SPLIT + GT covers only the route-planning questions whose landmarks the annotations identify.