Rebuilding Visual Spatial Intelligence Evaluation for Accurate Assessment of VLM 3D Reasoning

Yiming Zhang, Jiacheng Chen*, Jiaqi Tan, Yongsen Mao, Wenhu Chen, Angel X. Chang*

SFU SIMON FRASER UNIVERSITY · 香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY · UNIVERSITY OF WATERLOO · amii

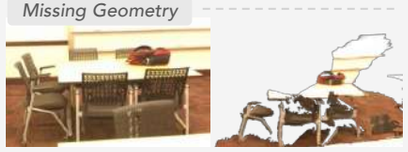
ReVSI — Rebuilding Visual Spatial Intelligence Evaluation

Hugging Face: 3dlg-hcvc/ReVSI · GitHub: 3dlg-hcvc/revsi · Visualization: revsi.site

Prior Spatial Intelligence Benchmarks Can Be Unreliable Due to Two Pitfalls!

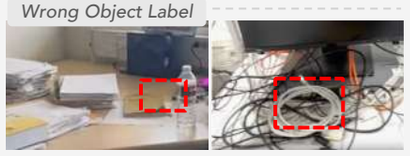
Pitfall 1: Bad Annotation Quality & 3D-to-Video Drift

Missing Geometry

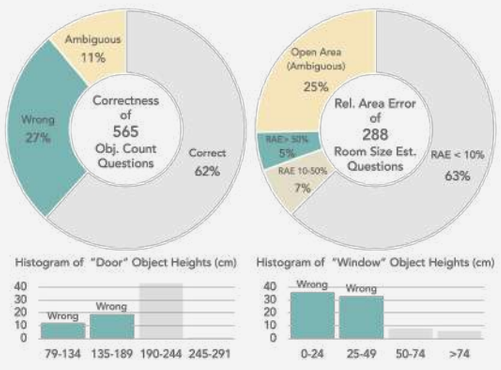


Chairs Count: 23 → 27

Wrong Object Label



Cup → Notebook Keyboard → Cables



VSI-Bench derives GT answers for question-answer pairs from low-quality 3D reconstructions and noisy annotations in scene datasets (e.g., ScanNet v2), leading to a substantial portion of incorrect GT answers across tasks.

Prior Benchmark: VSI-Bench

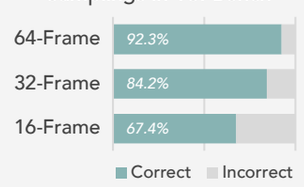
- What is the size of this room (m^2) ?
- How many chairs are in this room?
- What is the height of the stool, in cm?
- What is the distance between cabinet and piano (meters) ?
- …

Pitfall 2: Video Sampling Matters

How Many Pillows Are In This Room? *VSI-Bench GT: 6*

ReVSI GT: ALL-FRAME: 6 | 64-FRAME: 4 | 32-FRAME: 3 | 16-FRAME: 1

GT Correctness Across Frame Samplings in VSI-Bench

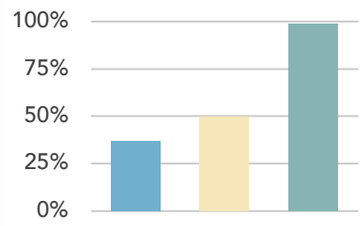


Prior evaluations assume full-scene access, whereas most vision-language models operate on sparsely sampled frames (e.g., 16-64). This mismatch leads to missing objects and geometries, rendering a large portion of questions unanswerable or incorrect.

VLMs Can Hallucinate

Gemini 3 Pro
Qwen3-VL-32B-Instruct
Cambrian-S-7B

Hallucination Rate on Dummy Video

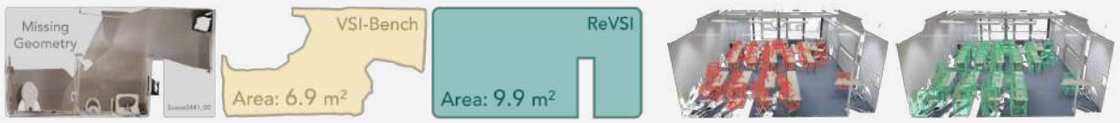


We probe models by masking queried objects from the video.

How ReVSI Address the Pitfalls?

High Quality - Expert-Level Annotations

Every ReVSI scene is is expert-annotated - without heuristic shortcuts - and passes multiple rounds of video-aware verification, ensuring accurate object names, physical size, and faithful room-area boundaries.



Room Size Annotations

Object Annotations (VSI-Bench vs ReVSI)

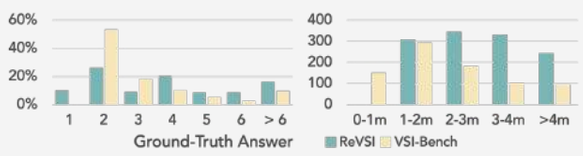
Great Diversity - Broader Scenes, Richer Objects

ReVSI covers more scenes, more objects, and larger object vocabulary than VSI-Bench, with open-vocabulary support.

	Scenes	Objects	Obj. Labels	Open-Vocab
VSI-Bench	288	3185	65	
ReVSI	**381**	**5365**	**504**	✓

Great Diversity - Broader Scenes, Richer Objects

More balanced answers reduce frequency-based guessing



Frame-Adaptive - What We Ask, What the Model Sees

ReVSI Ground-truth answers adapt to different input frame budget, reflecting what the model actually sees.

Question: How many chairs are in the room?

Answer:

All-Frame	64-Frame	32-Frame	16-Frame
12	11	9	6

Experiments & Findings

Method	Frames	Numerical Question				Multiple-Choice Question		
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
Baseline								
Chance (Random)	ALL	-	-	-	-	23.7 (25.0)	26.8 (36.1)	26.0 (28.3)
Chance (Frequency)	ALL	52.2 (62.1)	40.1 (32.0)	17.4 (29.9)	20.9 (33.1)	25.8 (25.1)	31.9 (47.9)	30.2 (28.4)
Proprietary Models (API)								
GPT-5.2	64	56.2 (57.1)	41.5 (33.4)	73.9 (64.6)	63.0 (59.0)	48.4 (48.0)	34.9 (33.3)	38.2 (36.7)
Gemini 3 Flash	1 FPS	65.7 (45.6)	53.1 (36.3)	77.6 (74.9)	52.8 (47.4)	64.6 (54.3)	47.9 (52.4)	41.8 (50.0)
Gemini 3 Pro	1 FPS	60.1 (45.3)	54.7 (38.3)	79.3 (73.0)	51.9 (47.4)	68.1 (70.0)	56.0 (60.8)	56.4 (65.3)
Open-Source Models								
Qwen3-VL-8B-Instruct	64	40.4 (70.0)	52.3 (50.5)	69.0 (74.7)	45.1 (63.3)	57.1 (57.3)	39.5 (52.3)	40.5 (33.5)
Qwen3-VL-32B-Instruct	64	46.9 (74.0)	65.0 (57.0)	70.4 (76.6)	55.8 (70.8)	53.8 (55.6)	34.0 (59.1)	47.3 (39.7)
InternVL3.5-8B	64	43.3 (72.7)	54.6 (40.3)	64.2 (68.4)	47.6 (65.3)	45.0 (57.0)	36.3 (48.6)	44.4 (35.6)
InternVL3.5-38B	64	43.8 (73.9)	60.6 (39.2)	70.2 (73.0)	58.4 (65.0)	57.4 (66.2)	45.9 (72.0)	42.7 (36.1)
LLaVA-Video-7B-Qwen2	64	31.3 (50.6)	1.4 (13.3)	52.5 (44.7)	16.7 (23.8)	38.3 (43.7)	33.3 (42.7)	38.4 (35.6)
LLaVA-Video-72B-Qwen2	64	40.1 (51.9)	29.6 (24.0)	59.3 (57.4)	27.9 (32.7)	39.6 (42.4)	24.8 (37.4)	43.0 (32.0)

Open-Source Models Are Systematically Overestimated

Method	Frames	Numerical Question				Multiple-Choice Question		
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
Qwen2.5-7B-Instruct+SigLIP2	–	–	–	–	–	–	–	–
Cambrian-S-7B	128	48.4 (73.2)	60.5 (50.5)	65.5 (74.9)	46.7 (72.2)	37.1 (71.1)	48.5 (76.2)	37.0 (41.8)
Qwen2.5-VL-7B-Instruct	4 FPS	36.9 (36.8)	15.0 (17.6)	49.7 (51.0)	29.0 (29.2)	31.5 (35.4)	29.5 (38.4)	36.7 (33.5)
VST-7B-SFT	4 FPS	35.4 (72.0)	52.6 (44.4)	67.9 (74.3)	47.2 (68.3)	49.2 (59.7)	36.9 (55.8)	35.4 (44.9)
Qwen2.5-VL-7B-Instruct	32	34.3 (43.7)	21.7 (22.3)	45.5 (49.2)	35.1 (37.5)	32.6 (40.1)	33.7 (38.9)	34.1 (32.0)
SpaceR-7B (SG-RLVR)	32	30.7 (61.9)	34.5 (28.6)	52.0 (60.9)	18.6 (35.2)	22.8 (38.2)	34.5 (46.0)	20.2 (31.4)
Qwen2.5-VL-3B-Instruct	16	18.7 (24.3)	15.6 (24.7)	16.8 (31.7)	– (22.6)	33.2 (38.3)	34.3 (41.6)	– (26.3)
Spatial-MLLM-4B-135k	16	40.7 (65.8)	45.3 (40.7)	46.8 (58.3)	– (55.6)	32.3 (43.2)	37.4 (55.5)	– (36.1)
Spatial-MLLM-4B-820k	16	41.5 (66.7)	40.0 (37.9)	53.1 (69.7)	– (55.7)	30.7 (52.0)	39.2 (54.9)	– (39.7)
LLaVA-Video-7B-Qwen2	32	29.9 (48.5)	1.5 (14.0)	53.0 (47.8)	19.3 (24.2)	39.1 (43.5)	33.8 (42.4)	38.8 (34.0)
VLM3R-7B	32	41.6 (70.2)	61.6 (49.4)	64.8 (69.2)	52.5 (67.1)	46.5 (65.4)	49.5 (80.5)	34.1 (45.4)

Spatial Fine-Tuning Does Not Reliably Generalize