Struct2D: A Perception-Guided Framework for Spatial Reasoning in MLLMs

Fangrui Zhu[1*], Hanhui Wang[1,3*], Yiming Xie[1], Jing Gu[4], Tianye Ding[1], Jianwei Yang[2], Huaizu Jiang[1]

[1]Northeastern University, [2]Microsoft Research, [3]University of Southern California, [4]University of California, Santa Cruz

Paper Code

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Can MLLMs reason about 3D space using only structured 2D representations derived from perception?

Contributions

- We propose a perception-guided 2D prompting strategy, **Struct2D** Prompting, and conduct a detailed **zero-shot analysis** that reveals MLLM's ability to perform 3D spatial reasoning from structured 2D input alone.
- We introduce **Struct2D-Set**, a large-scale instructional tuning dataset with automatically generated, fine-grained QA pairs covering eight spatial reasoning categories grounded in 3D scenes.
- We fine-tune an open-source MLLM (Qwen2.5-VL) on our Struct2D-Set to achieve competitive performance across several spatial reasoning benchmarks, validating the real-world applicability of our framework.

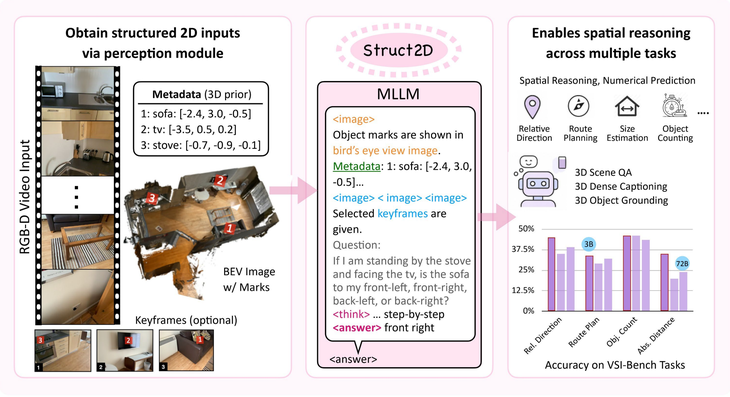

Figure 1. Overview of our **Struct2D** framework for enabling spatial reasoning in multimodal large language models (MLLMs).

Struct2D Prompting

- **Struct2D** transforms RGB-D videos into structured 2D inputs—a bird's eye view (***BEV***) image with *filtered object* ***marks***, *object-centric* ***metadata***, and optionally *object-focused* ***keyframes*** (as shown in Fig. 1).
- We evaluate Struct2D on a VSI-Bench subset, comparing with prior 2D and video-based prompting. With *noisy* object detections, Struct2D outperforms both baselines across all major spatial reasoning tasks while using only one input image. (Tab. 1)
- When provided with *ground-truth* object detections, Struct2D achieves **near-human performance** across most spatial reasoning tasks. (Tab. 1)

Struct2D-Set Construction

- Built from ScanNet, ScanNet++, and ARKitScenes datasets.
- Consists of 200K QA pairs across 8 spatial reasoning tasks.

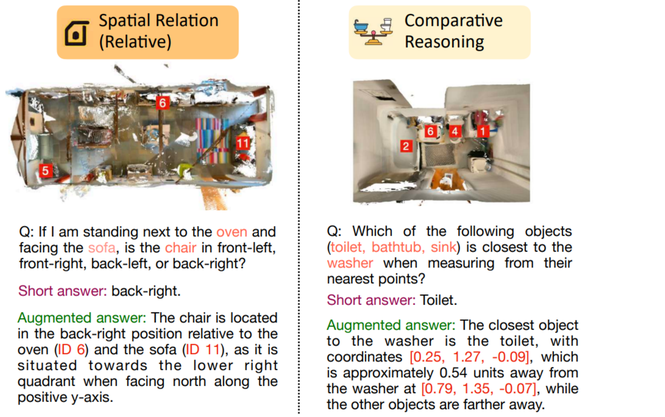

Figure 2. QA examples of Struct2D-Set. For complex spatial reasoning tasks, QA pairs include both concise geometric answers and detailed reasoning explanations.

Experiment Results Highlights

Table 1. Zero-shot evaluation of GPT-o3 on the VSI-Bench subset. We compare our Struct2D approach against baselines using 16 video frames and GPT4Scene's prompting solution. Struct2D achieves superior performance across most tasks.

Settings	# images	Cost ($)	Avg.	Numerical Answer			Multiple-Choice Answer		
				Obj. Count	Abs. Dist.	Room Size	Rel. Dist.	Rel. Dir.	Route Plan
VSI-Bench [81]	16	105.07	48.6	44.3	34.1	50.9	51.0	49.4	61.9
GPT4Scene [62]	9	78.67	50.3	51.5	35.3	**58.0**	50.5	47.9	58.8
Ours (Noisy Objects)	1	**27.25**	**56.1**	**52.8**	**38.4**	48.9	**60.0**	**60.1**	**76.2**
Ours (GT Objects)	1	**27.25**	83.8	93.8	90.6	47.4	96.5	94.4	80.1

Table 2. Performance comparison of various models on VSI-Bench. Our model fine-tuned with Struct2D-Set surpasses both the zero-shot Struct2D prompting and video-based tuning baseline.

Methods	Avg.	Numerical Answer				Multiple-Choice Answer		
		Obj. Count	Abs. Dist.	Room Size	Obj. Size	Rel. Dist.	Rel. Dir.	Route Plan
Open-source Models								
InternVL2-2B [15]	30.3	21.8	24.9	35.0	22.0	33.8	44.2	30.5
InternVL2-8B [15]	33.9	23.1	28.7	39.8	48.2	36.7	30.7	29.9
LongVILA-8B [79]	21.1	29.1	9.1	0.0	16.7	29.6	30.7	32.5
VILA-1.5-8B [45]	29.5	17.4	21.8	18.8	50.3	32.1	34.8	31.0
LongVA-7B [90]	31.1	38.0	16.6	22.2	38.9	33.1	43.3	25.4
LLaVA-NeXT-Video-7B [95]	36.3	**48.5**	14.0	24.2	47.8	**43.5**	42.4	34.0
LLaVA-OneVision-0.5B [39]	31.2	46.1	28.4	28.3	15.4	28.9	36.9	34.5
LLaVA-OneVision-7B [39]	33.5	47.7	20.2	12.3	47.4	42.5	35.2	29.4
R1-Zero-VSI [44] (Qwen2-VL-7B)	32.1	39.4	25.0	43.2	25.8	32.6	30.9	27.8
R1-Zero-VSI [44] (Qwen2-VL-7B) + SFT	38.8	44.7	27.6	**50.4**	46.1	34.0	35.7	33.0
Ours								
Qwen2.5-VL-3B	25.6	27.0	22.0	25.6	32.5	17.5	28.9	25.6
Qwen2.5-VL-3B (Struct2D Prompting)	29.4	46.6	24.6	22.3	33.6	21.2	30.5	27.2
Qwen2.5-VL-3B (Baseline)	33.9	24.6	34.0	46.4	53.5	21.2	30.5	27.2
Qwen2.5-VL-3B (SFT)	**41.9**	46.0	**34.7**	42.6	**56.4**	35.1	**44.9**	**33.5**

Table 3. 3D grounding results on ScanRefer.

Methods	ScanRefer (val)	
	Acc@0.25	Acc@0.50
Task-Specific Model		
3DVG-Transformer [96]	47.6	34.7
3DJCG [6]	49.6	37.3
D3Net [11]	–	37.9
M3DRef-CLIP [94]	51.9	44.7
3D LLM Based Model		
Chat-Scene [30]	55.5	50.2
Vision LLM Based Model		
Qwen2-VL-7B [71]	5.4	5.1
Qwen2-VL-7B (GPT4Scene [62])	40.5	36.7
Qwen2.5-VL-7B (Ours)	51.7	48.5

Question: Paint a picture of the visual attributes of <OBJ11>, then explore how it relates spatially to other elements in the scene.

Ground Truth: This is a brown table. It is by the couch.

Ours: The table is to the right of the couch. The table is a brown rectangle.

Figure 3. 3D dense captioning output example.

Contact: {zhu.fang, wang.hanh, xie.yim, ding.tian, h.jiang}@northeastern.edu, jw2.yang@gmail.com, hanhuiwa@usc.edu, jgu110@ucsc.edu