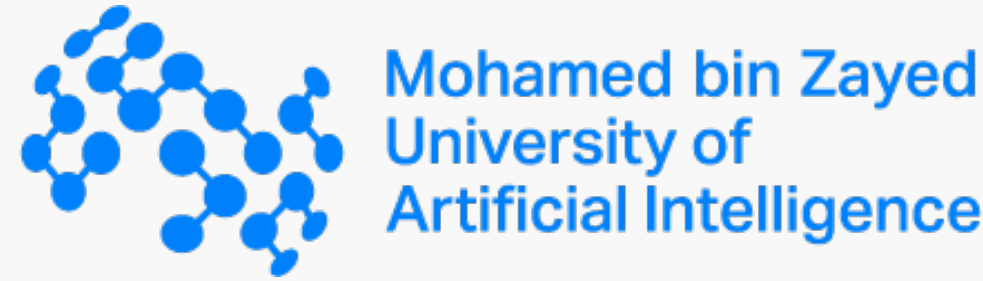


SURPRISE3D: A Dataset for Spatial Understanding and Reasoning in Complex 3D Scenes

Jiaxin Huang^{*1}, Ziwen Li^{*1}, Hanlve Zhang^{*1}, Runnan Chen², Zhengqing Gao¹, Xiao He³, Yandong Guo³, Wenping Wang⁴, Tongliang Liu^{1,2}, Mingming Gong^{†1,5}

¹ Mohamed Bin Zayed University of Artificial Intelligence, ² The University of Sydney, ³ AI2Robotic, ⁴ Texas A&M University, ⁵ The University of Melbourne



Motivation

- Embodied AI requires understanding the relationship between language and 3D spatial context, yet current datasets rely heavily on semantic cues (object names) rather than true spatial reasoning.
- Existing 3D vision-language models often exploit shortcuts (category matching) instead of reasoning about geometry.
- SURPRISE3D fills this gap by introducing a dataset focused on language-guided spatial reasoning segmentation.**

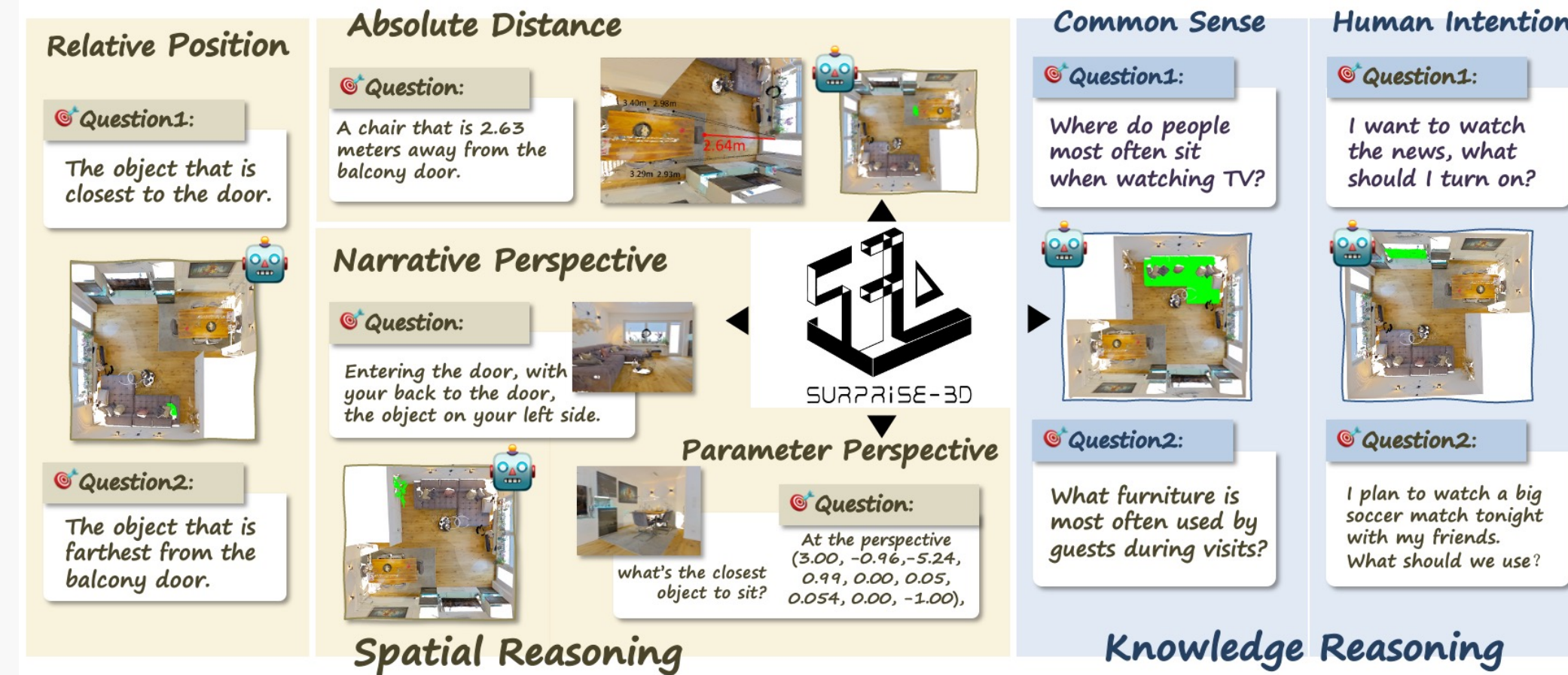
Target Problem

- Existing 3D vision-language datasets often entangle semantic recognition with spatial reasoning, allowing models to succeed by exploiting object names rather than truly understanding spatial context.
- This prevents robust evaluation of reasoning skills such as viewpoint-dependent references or distance-based relations.
- The key challenge is to create a dataset that isolates spatial understanding from semantics and tests genuine reasoning capabilities across complex 3D layouts.
- SURPRISE3D addresses this gap by providing large-scale, human-verified spatial queries and a standardized benchmark to rigorously evaluate 3D spatial reasoning segmentation.**

Dataset	Output	Multi-target	Lang. Source	Spatial Reason	1st observ.	Shortcut Free
CLEVR3D [42]	Lang	-	Template	✗	-	-
Scan2Cap [7]	Lang	-	Human	✗	-	-
ScanQA [4]	Lang	✗	Human	✗	-	-
3DVQA [11]	Lang	✗	Human	✗	-	-
SQA3D [35]	Lang	-	Human	✓	✓	✗
ScanScribe [52]	Lang	✗	Template,LLM	-	-	-
3DMV-VQA [14]	Lang	-	Template	✗	✗	✗
M3DBench [27]	Lang	✓	LLM	✗	✗	✗
SceneVerse [22]	Lang	✗	Template,LLM	✗	✗	-
MSQA [29]	Lang	-	Human,LLM	✓	✓	✗
VLA-3D [48]	Lang	✓	Template	✓	✓	✗
ExCap3D [45]	Lang	✓	LLM	✗	✓	✗
ReferIt3D [3]	BBox	✓	Human,Template	✗	✓	✗
ScanRefer [6]	BBox	✗	Human	✗	✓	✗
3D-DenseOG [21]	BBox	✓	Human	✗	✓	✗
Grounded 3D-LLM [10]	Mask	✗	Template,LLM	✓	✗	✗
ScanEnts3D [11]	BBox	✓	Human	✓	✗	✗
PhraseRefer [46]	BBox	✓	Human	✗	✗	✗
EmbodiedScan [39]	BBox	✓	Template,LLM	✓	✓	✗
3D-LLM [15]	Lang + BBox	✓	LLM	✓	✗	✗
LL3DA [8]	Lang + BBox	✓	Template,LLM	✗	✗	✗
3DMIT [28]	Lang + BBox	✗	LLM	✗	✓	✗
3D-GRAND [43]	Lang + BBox	✓	Template,LLM	✗	✓	✗
Segpoint [13]	Mask	✗	LLM	✓	✗	✓
ScanReason [51]	BBox	✗	LLM	✓	✗	✓
Reason3D [19]	Mask	✗	-	✓	✗	✓
Intent3D [24]	BBox	✓	LLM	✗	✗	✓
SURPRISE3D (Ours)	Mask	✓	Human,LLM	✓	✓	✓

Task & Benchmark

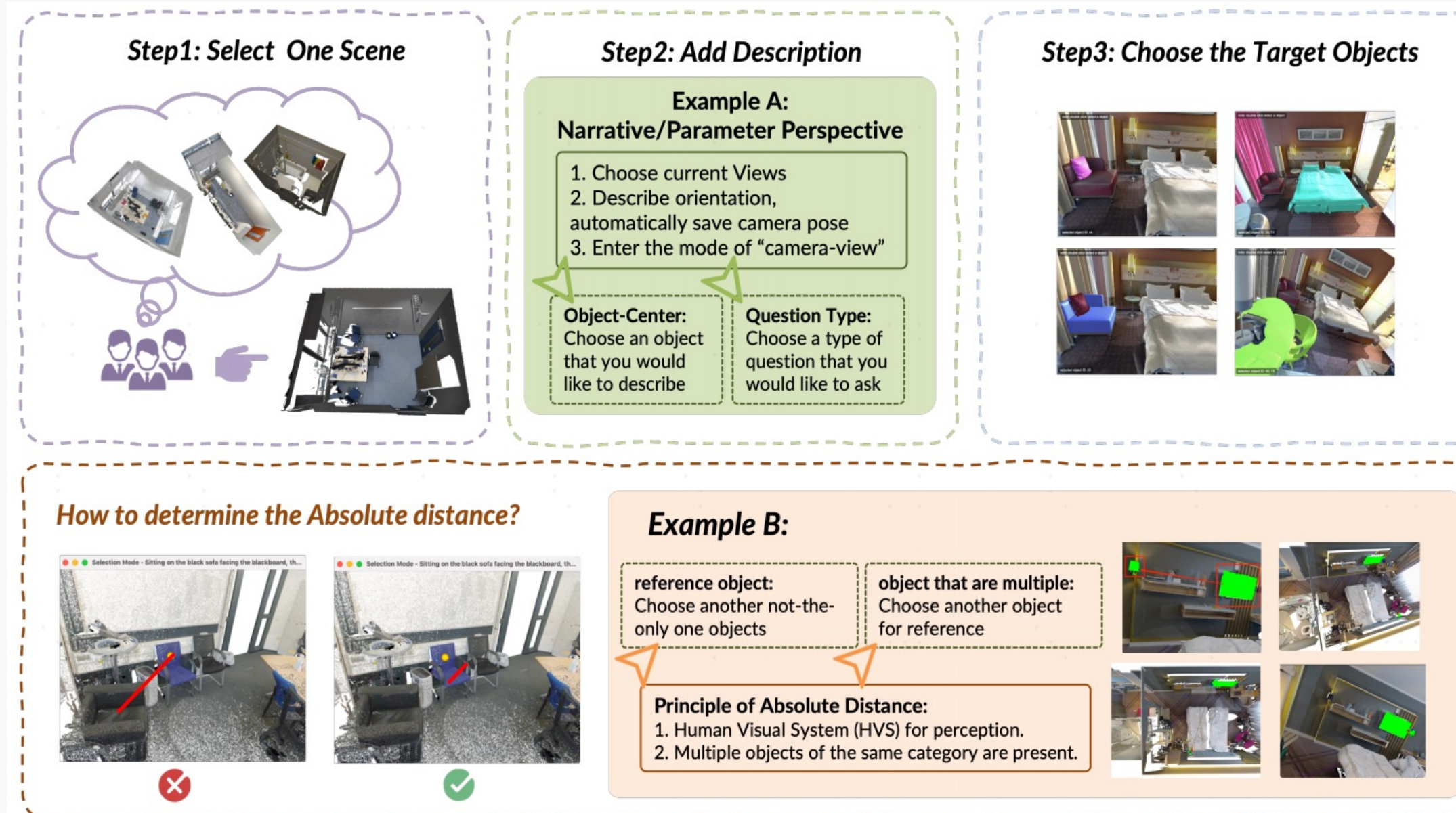
- 3D Spatial Reasoning Segmentation (3D-SRS)** aims to predict a 3D mask for objects in a scene that satisfy a language query describing spatial relations.
- It covers **four reasoning types: relative position, narrative perspective, parametric perspective, and absolute distance.**
- Evaluation uses mIoU, precision, and recall at IoU 0.5/0.75.
- Scenes are split to avoid overlap, with a leaderboard ensuring fair comparison.



Annotation Pipeline

We develop two annotation pipelines for different reasoning types:

- Spatial reasoning**, human annotators fix a camera viewpoint, write spatial queries (e.g., “the chair to my right”), and mark the target mask in 3D, ensuring consistent egocentric interpretation.
- Knowledge reasoning**, an LLM-human hybrid pipeline generates and verifies question-answer pairs, combining scalability with human quality control.



Experiments & Results

Challenging for SOTA 3D grounding & 3D-LLMs in zero-shot. Fine-tuning on SURPRISE3D yields large gains across reasoning types. Takeaway: models still lack true 3D spatial reasoning → our benchmark fills the gap.

Task Type	Metric	Method					
		MLLMfor3D	3D-Vista	Reason3D	Intent3D	ChatScene	Avarage
Knowledge Reasoning							
Common-sense	A25	20.42	6.14	6.97	10.01	7.86	10.28
	A50	13.42	6.14	3.11	3.24	4.01	5.98
	mIoU	12.75	—	4.79	—	—	8.74
Human Intention	A25	22.38	8.26	11.33	15.84	1.64	11.89
	A50	13.62	8.26	6.03	5.82	1.00	7.48
	mIoU	11.91	—	7.51	—	—	6.31
Spatial Reasoning							
Narrative Perspective	A25	15.07	6.50	8.39	2.65	0.00	6.52
	A50	13.62	6.50	4.88	0.77	0.00	4.85
	mIoU	11.56	—	5.63	—	—	2.75
Parametric Perspective	A25	4.25	3.65	7.91	2.62	0.00	3.29
	A50	3.20	3.65	4.82	0.79	0.00	2.66
	mIoU	2.93	—	5.68	—	—	2.31
Relative Position	A25	7.78	6.52	9.57	4.30	1.38	5.99
	A50	5.81	6.52	6.38	0.97	0.00	4.47
	mIoU	4.92	—	6.78	—	—	3.86
Absolute Distance	A25	11.90	7.61	9.10	2.41	1.39	6.48
	A50	10.62	7.61	2.60	0.74	0.00	4.06
	mIoU	9.24	—	5.25	—	—	3.47
Overall Reasoning							
Overall	A25	13.63	6.40	9.09	8.63	3.59	8.27
	A50	10.05	6.40	4.57	2.94	1.77	4.91
	mIoU	8.89	—	6.08	—	—	4.72

The results of various methods on different reasoning tasks under **zero-shot** setting

Task Type	Metric	Method					
		MLLMfor3D	3D-Vista	Reason3D	Intent3D	ChatScene	Avarage
Knowledge Reasoning							
Common-sense	A25	25.41	19.36	18.08	30.09	13.56	21.30
	A50	23.92	19.36	8.97	15.22	4.37	14.37
	mIoU	19.40	—	11.92	—	—	13.39
Human Intention	A25	28.42	22.36	17.98	31.16	13.80	22.74
	A50	24.51	22.36	10.81	18.08	4.70	17.29
	mIoU	19.47	—	11.98	—	—	15.73
Spatial Reasoning							
Narrative Perspective	A25	22.38	25.77	11.30	24.91	13.98	19.67
	A50	20.40	25.77	7.71	15.58	4.28	14.75
	mIoU	18.44	—	9.03	—	—	13.74
Parametric Perspective	A25	10.04	3.87	11.52	12.57	4.58	8.52
	A50	9.33	3.87	7.59	6.13	2.36	5.86
	mIoU	7.50	—	8.85	—	—	8.18
Relative Position	A25	22.61	23.86	11.51	12.90	12.48	16.67
	A50	18.76	23.86	8.18	7.49	1.39	12.46
	mIoU	14.70	—	8.88	—	—	12.79
Absolute Distance	A25	25.30	18.92	12.80	5.75	7.42	14.04
	A50	20.37	18.92	4.45	1.86	3.71	10.66
	mIoU	19.23	—	8.16	—	—	13.70
Overall Reasoning							
Overall	A25	22.36	18.34	16.14	23.98	11.58	18.48
	A50	19.55	18.34	9.06	13.10	3.99	13.00
	mIoU	16.46	—	11.00	—	—	13.73

The results of various methods on different reasoning tasks under **fine-tuned** setting



First-author Contact

Github

Project Page

Arxiv