ExHall D Poster #355
Sat 14 Jun 10:30 -12:30 CDT

PerLA

Perceptive 3D Language Assistant

Guofeng Mei[1] Wei Lin[2] Luigi Riz[1] Yujiao Wu[3] Fabio Poiesi[1] Yiming Wang[1]

[1]Fondazione Bruno Kessler, Italy [2]JKU Linz, Austria [3]CSIRO, Australia

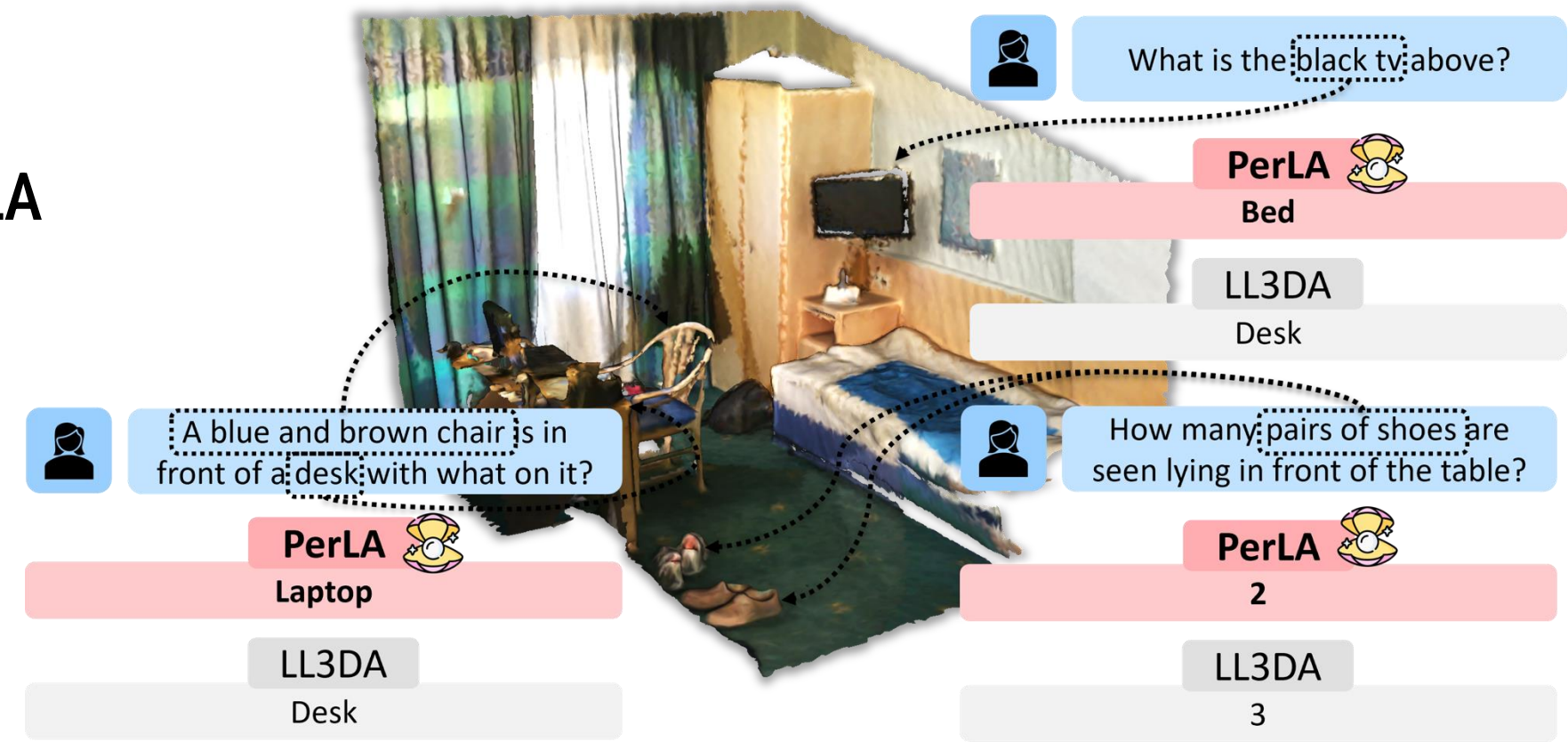
FONDAZIONE BRUNO KESSLER JOHANNES KEPLER UNIVERSITÄT LINZ CSIRO

CVPR Nashville JUNE 2025

Agenda

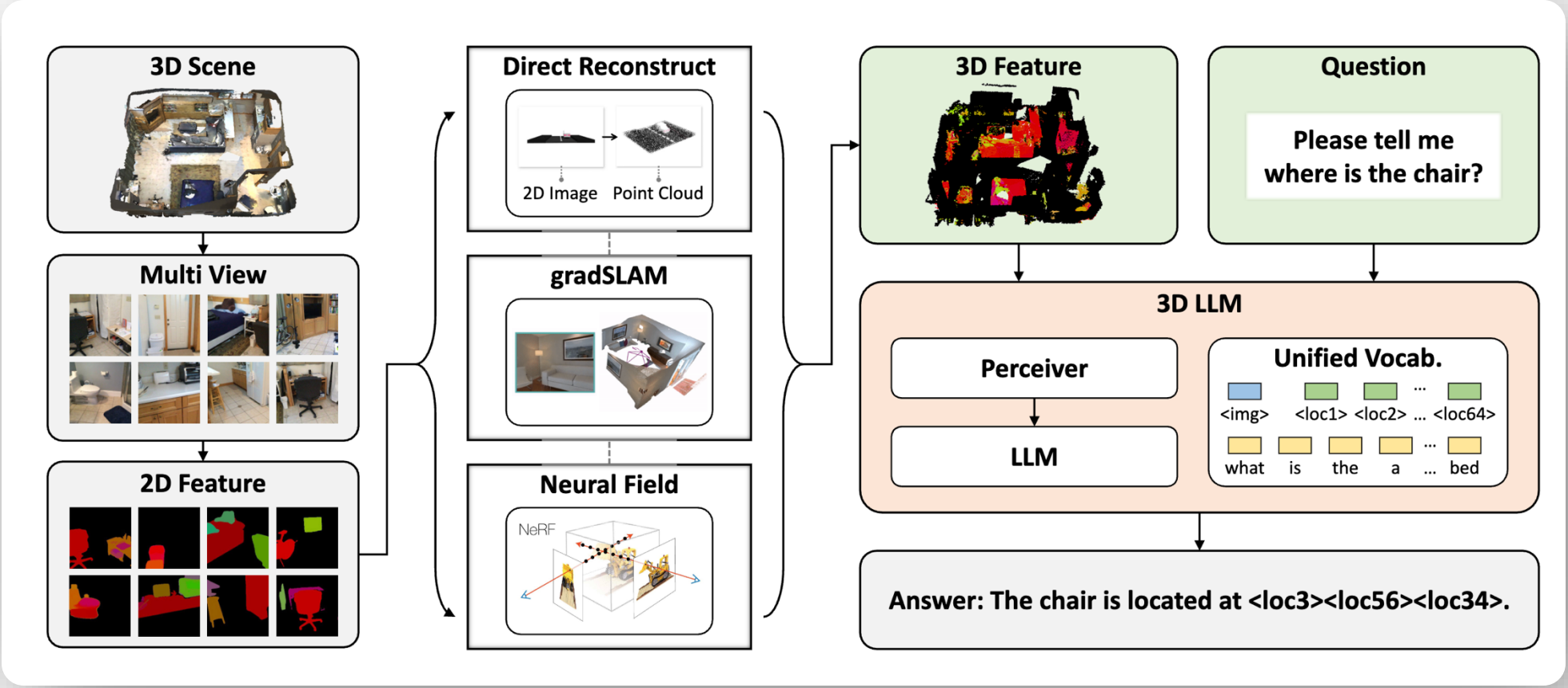
PerLA

- Introduction
- Motivation
- Solution: PerLA
- Results
- Conclusions



Introduction

2D encoder: multi-view image input (3DLLM-style)

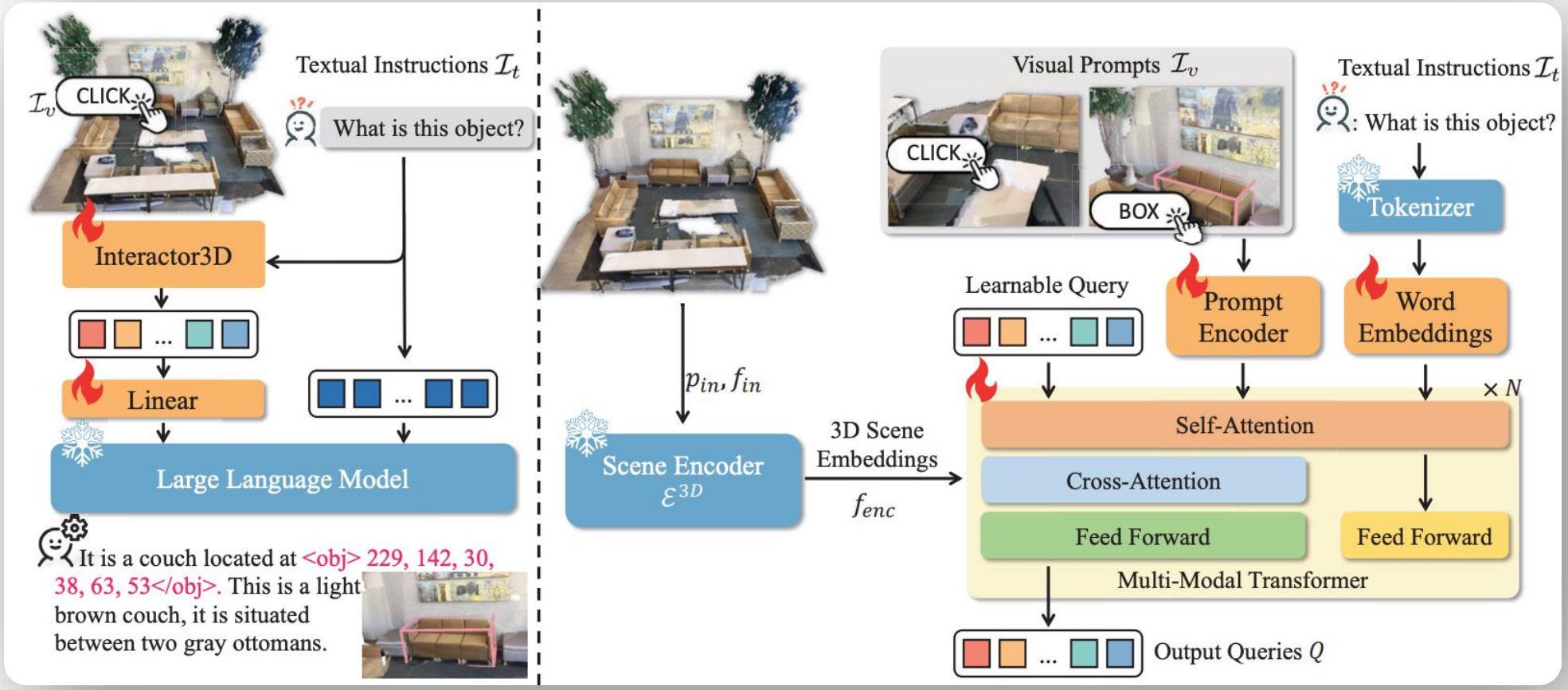


Multi-view inputs: high computational cost, low efficiency, and poor 3D spatial fidelity

Hong, Yining, et al. 3d-llm: Injecting the 3d world into large language models. NeurIPS, 2023.

Introduction

3D encoder: 3D point input (LL3DA-style)



3D inputs: fast and spatially accurate, but sampling degrades performance for high-res point clouds

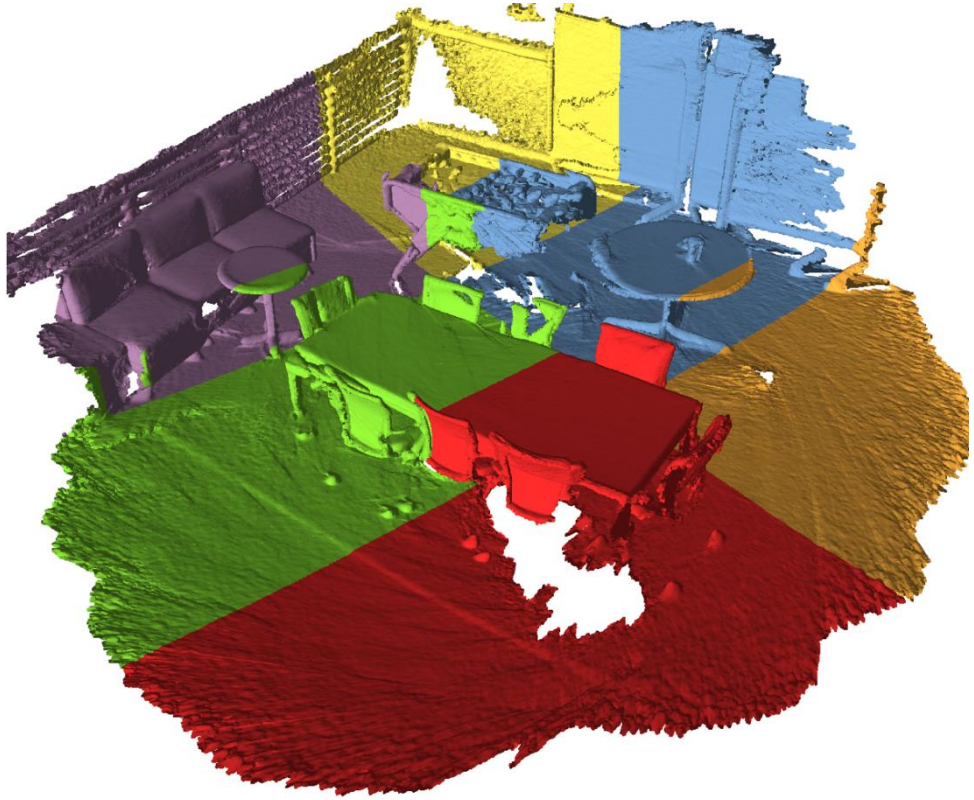
Chen, Sijin, et al. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. CVPR. 2024.

Motivation

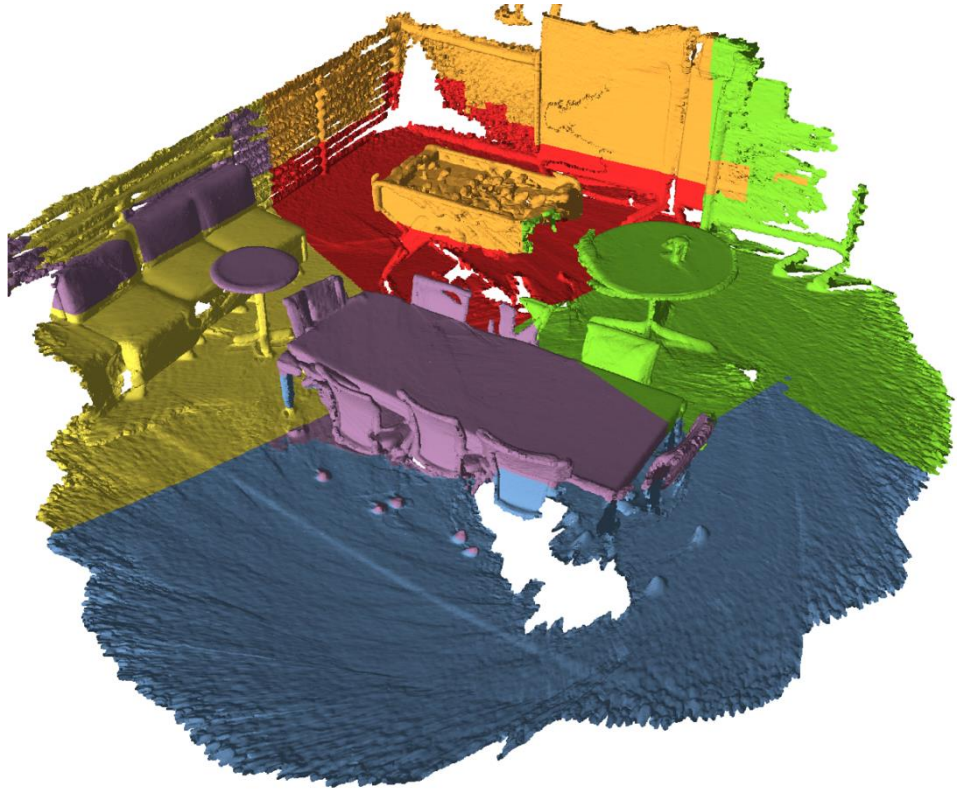
Partition point cloud into parts: **How?**

PerLA

Naive partition



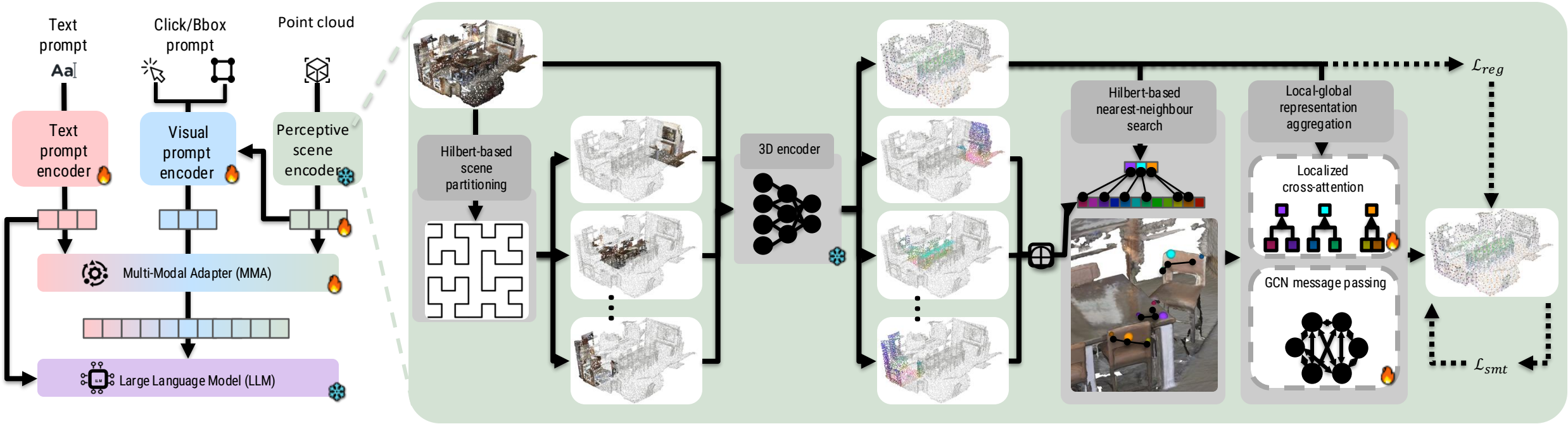
Hilbert partition



Geometric and semantic consistent

Solution

Our framework



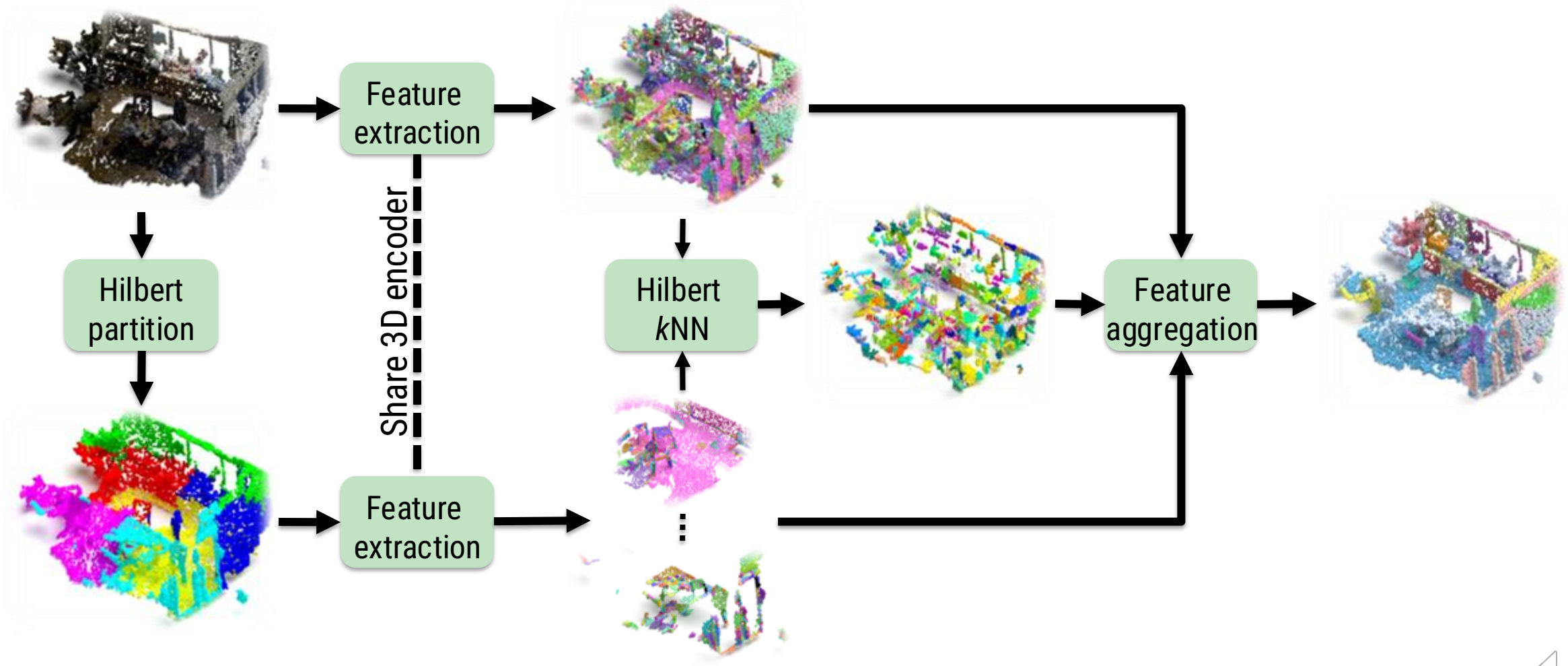
Hilbert-based scene partitioning → Global and partial scene encoding → Hilbert-based nearest-neighbor search → Local-global representation aggregation

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Solution

PerLA

From start to finish: a layered view of our approach



Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Results

3D Question Answering on ScanQA

Method	Type	ScanQA validation				ScanQA test w/ object				ScanQA test w/o object			
		C ↑	B4 ↑	M ↑	R ↑	C ↑	B4 ↑	M ↑	R ↑	C ↑	B4 ↑	M ↑	R ↑
ScanQA	CLS	64,86	10,08	13,14	33,33	67,29	12,04	13,55	34,34	60,24	10,75	12,59	31,09
Clip-Guided	-	-	-	-	-	69,53	14,64	13,94	35,15	62,83	11,73	13,28	32,41
Multi-CLIP	CLS	-	-	-	-	68,70	12,65	13,97	35,46	63,20	12,87	13,36	32,61
3D-VLP	CLS	66,97	11,15	13,5	34,51	70,18	11,23	14,16	35,97	63,40	**15,84**	13,13	31,79
3D-VisTA	-	-	-	-	-	68,60	10,50	13,80	35,50	55,70	8,70	11,69	29,60
3D-LLM	GEN	69,40	12,00	14,50	35,70	69,60	11,60	14,90	35,30	-	-	-	-
LL3DA	GEN	76,79	13,53	15,88	37,31	78,16	13,97	16,38	38,15	70,29	12,19	14,85	35,17
LL3DA (repr)	GEN	74,37	13,50	15,09	36,31	-	-	-	-	-	-	-	-
PerLA	GEN	**78,13**	**14,49**	**17,44**	**39,60**	**80,91**	**17,21**	**16,49**	**40,71**	**74,82**	14,97	**15,23**	**38,18**
Δ LL3DA	-	+1,34	+0,96	+1,56	+2,29	+2,75	+3,24	+0,11	+2,56	+4,53	+2,78	+0,38	+3,01

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Results

3D Dense Captioning on ScanRefer and Nr3D

PerLA

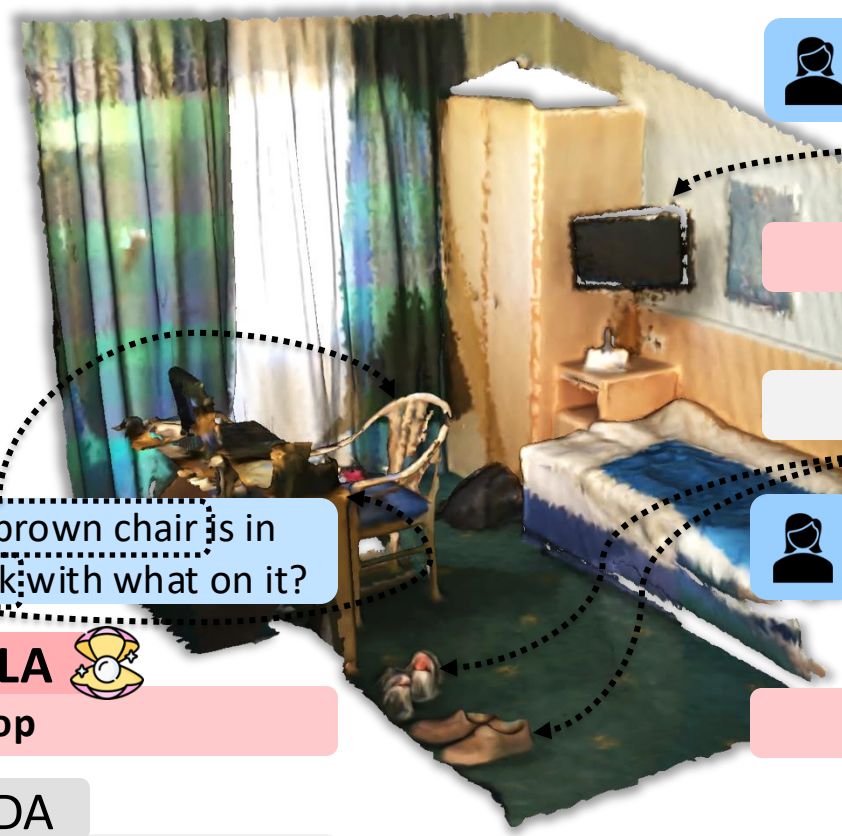
Method	ScanRefer@0.25				ScanRefer@0.5				Nr3D@0.5			
	C ↑	B4 ↑	M ↑	R ↑	C ↑	B4 ↑	M ↑	R ↑	C ↑	B4 ↑	M ↑	R ↑
Scan2Cap	56,82	34,18	26,29	55,27	39,08	23,32	21,97	44,78	27,47	17,24	21,80	49,06
MORE	62,91	36,25	26,75	56,33	40,94	22,93	21,66	44,42	-	-	-	-
REMAN	62,01	36,37	26,76	56,25	45,00	26,31	23,13	46,96	34,81	20,37	22,71	50,90
3DJCG	64,70	40,17	27,63	59,23	49,48	31,63	24,36	50,80	38,06	22,82	23,77	52,99
3DVLP	70,73	41,03	28,14	**59,72**	54,94	32,31	24,83	51,51	-	-	-	-
Vote2Cap	71,45	39,34	28,25	59,63	61,81	34,46	26,22	54,40	43,84	26,68	25,41	54,43
LL3DA	74,17	41,41	27,76	59,53	65,19	36,79	25,97	55,06	51,18	28,75	25,91	56,61
PerLA	**77,92**	**43,41**	**28,97**	59,69	**69,41**	**38,02**	**29,07**	**56,80**	**55,06**	**31,24**	**28,52**	**59,13**
Δ LL3DA	+3,75	+2,00	+1,21	+0,16	+4,22	+2,30	+2,27	+1,74	+3,88	+2,49	+2,61	+2,52

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Results

Qualitative example on 3D Question Answering

PerLA



What is the black tv above?

PerLA

Bed

LL3DA

Desk

A blue and brown chair is in front of a desk with what on it?

PerLA

Laptop

LL3DA

Desk

How many pairs of shoes are seen lying in front of the table?

PerLA

2

LL3DA

3

Improved Spatial Reasoning

Enhanced Detail Recognition

Contextual Understanding

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Results

Qualitative example on 3D Dense Captioning

PerLA

Describe this [bbox] object.

PerLA

This is a round table.
It is in the center of the room.

LL3DA

There is a rectangular table.
It is to the left of the couch.

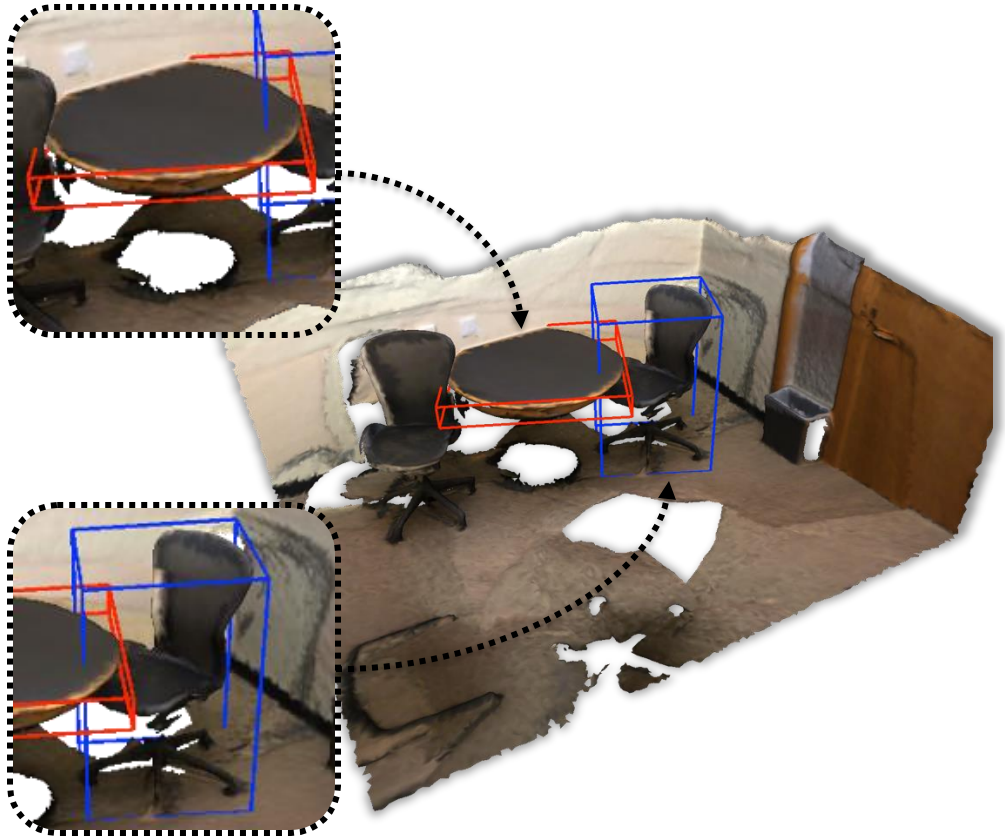
Describe this [bbox] object.

PerLA

This is a black chair.
It is to the right of another chair.

LL3DA

This is a brown chair.
It is to the right of the door.



Improved Spatial Reasoning

Enhanced Detail Recognition

Contextual Understanding

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Results

Ablation study results on the ScanQA benchmark

Method	ScanQA			
	C ↑	B4 ↑	M ↑	R ↑
LL3DA†	74,54	12,89	15,11	36,96
Global	74,49	13,50	15,16	36,55
Local	73,55	12,98	14,95	36,07
w/o GCN	77,61	13,87	15,93	39,28
PerLA	**78,13**	**14,49**	**17,44**	**39,60**

Effect of different aggregation strategies

Parts	ScanQA			
	C ↑	B4 ↑	M ↑	R ↑
4	76,9	13,6	16,1	37,3
6	**78,1**	**14,5**	**17,4**	**39,6**
8	78,0	14,5	17,2	39,4

Effect of part partitioning

78.13
PerLA Score
Highest ScanQA CIDEr score

3.59%
Improvement
Over LL3DA baseline

6
Optimal Parts
Best partition count

Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., & Wang, Y. . PerLA: Perceptive 3D language assistant. CVPR 2025.

Conclusion

Our contributions

PerLA

Point cloud is **serialized** via a **Hilbert curve** and **evenly partitioned** to adapt resolution to scene complexity

Global and **local features** are extracted with a **shared 3D encoder**, capturing scene context and local details

Hilbert-based indexing enables efficient, geometry-aware matching between global and local superpoints

Features are aggregated by localized cross-attention and refined via GCN message passing

ExHall D Poster #355
Sat 14 Jun 10:30 -12:30 CDT

PerLA

Perceptive 3D Language Assistant

Thank you for your attention!