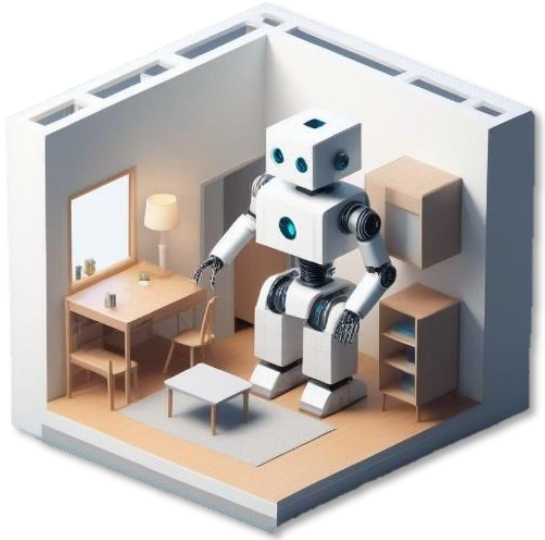
面向三维场景理解的视觉与语言模型研究

Thesis Defense for “Research on Vision and Language Models for 3D Scene Understanding”

陈思锦
21210720102



指导老师：陈涛

目录
Contents

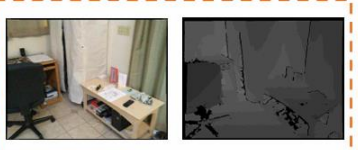
研究背景与意义

Research Background

研究背景与意义——相关背景：三维数据与表征

三维数据表征形式多样

(a) 三维点云

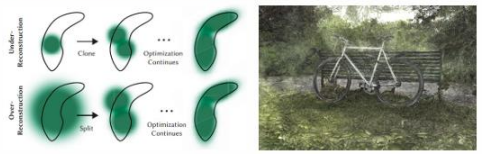


(c) 深度图

(d) 三维体素

(e) 三维网格

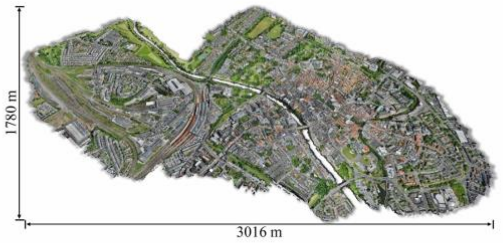
(b) 多视图图像



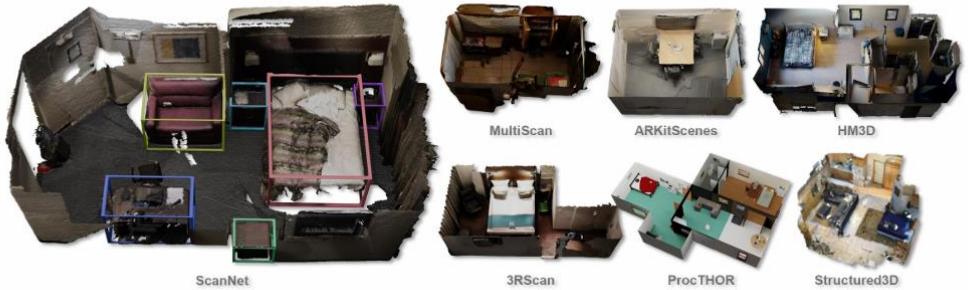
(f) 高斯溅射

三维数据尺度差异明显

(a) 三维物体数据集



(b) 三维城市数据集



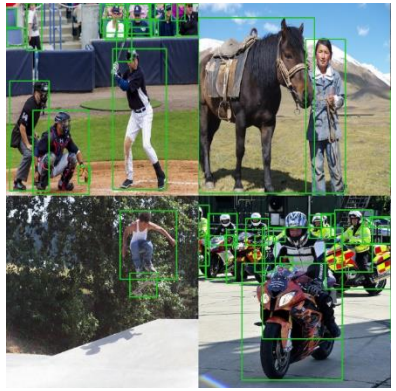
(c) 三维场景数据集

研究背景与意义——相关背景：视觉与语言

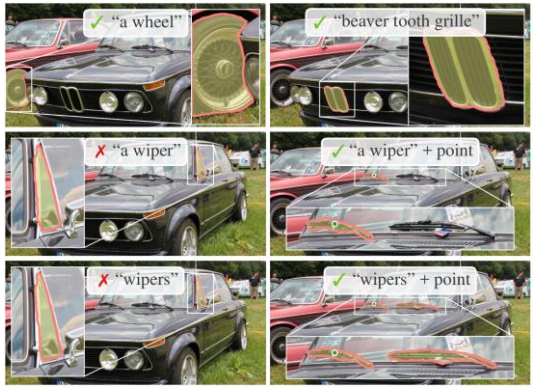
计算机视觉任务逐渐复杂，不局限于有限的类别标签

图像分类

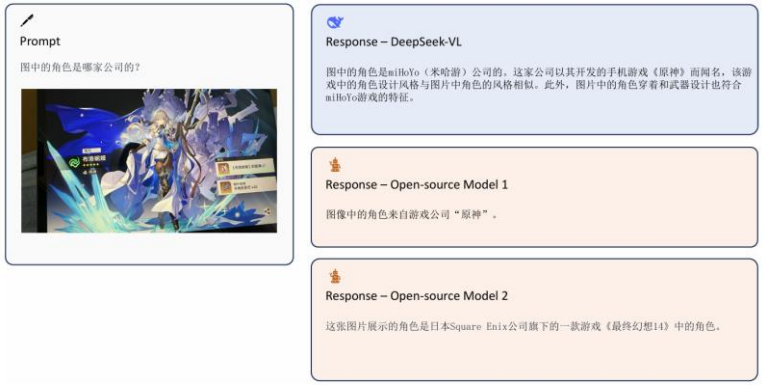
图像检测



开放词汇



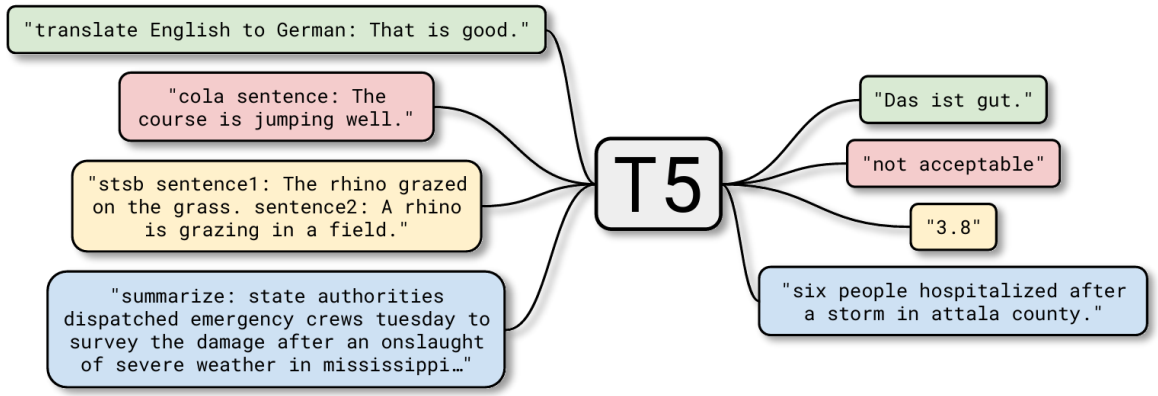
视觉推理



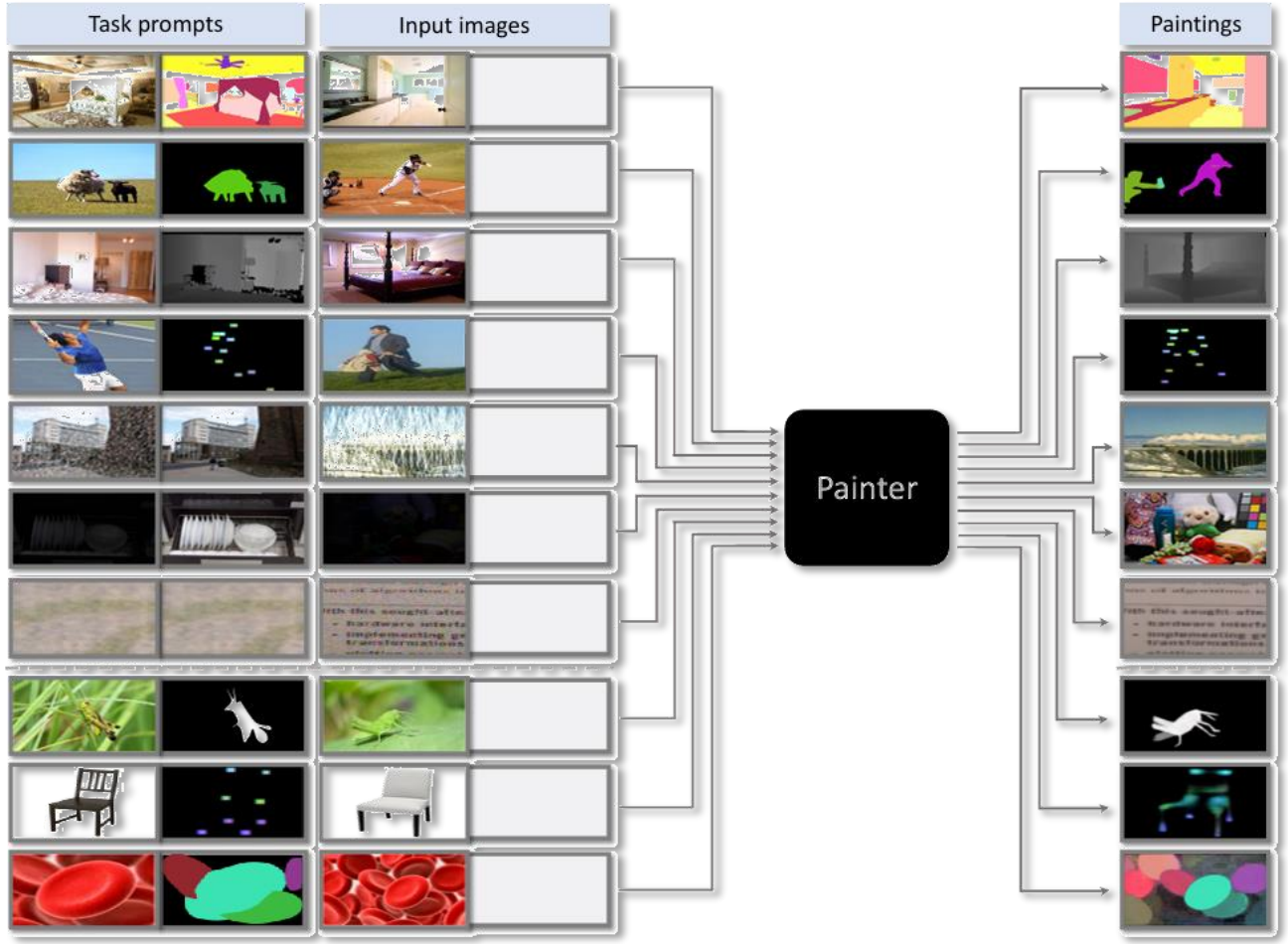
闭集视觉识别、定位任务

+ 自然语言：高阶抽象的语义信息

研究背景与意义——相关背景：构建通才基础模型



图片来源：https://arxiv.org/pdf/1910.10683

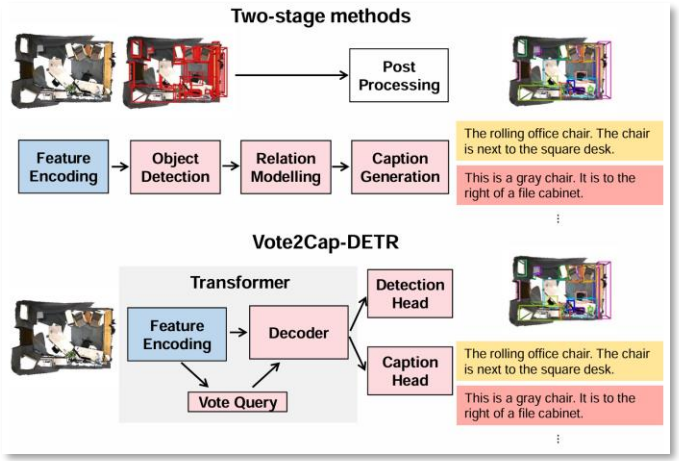


图片来源：https://arxiv.org/abs/2212.02499

研究背景与意义——Overview

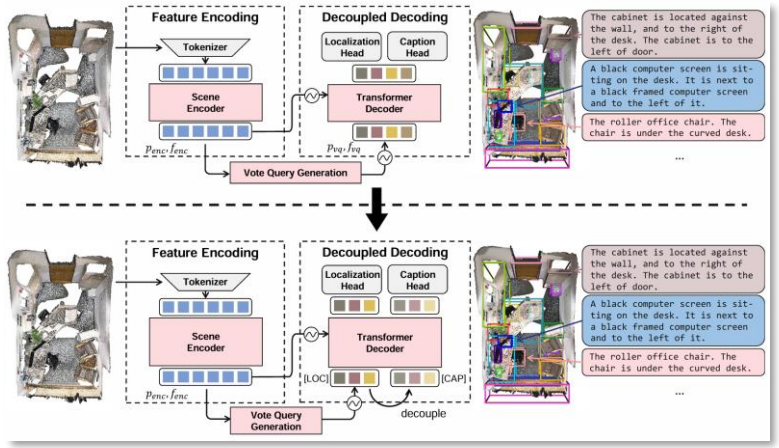
#Topic 1：用语言进行更好三维理解

三维场景感知瓶颈



Vote2Cap-DETR [CVPR 2023]

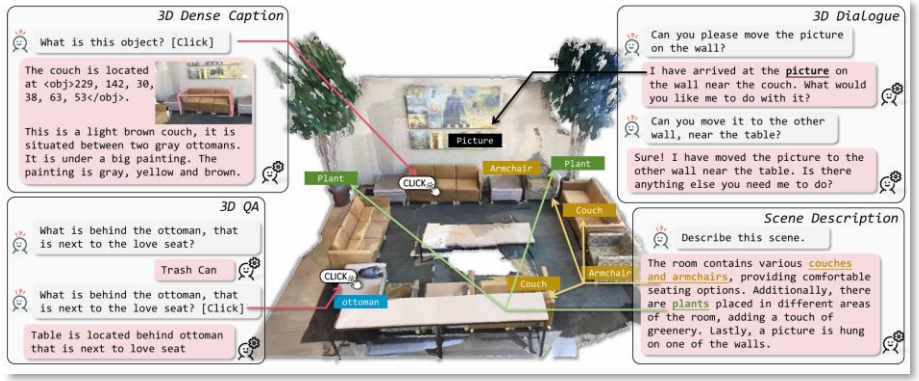
子任务间特征表征瓶颈



Vote2Cap-DETR++ [T-PAMI 2024]

#Topic 2：大语言三维通才模型

不同任务间的互相促进



LL3DA [CVPR 2024]

基于Transformer的端到端三维场景目标定位与描述生成

End-to-End 3D Dense Captioning with Vote2Cap-DETR, [CVPR 2023]

基于Transformer的端到端三维场景目标定位与描述生成

输入：三维场景

输出：三维目标框以及对应文本描述

Figure taken from Scan2Cap [1]

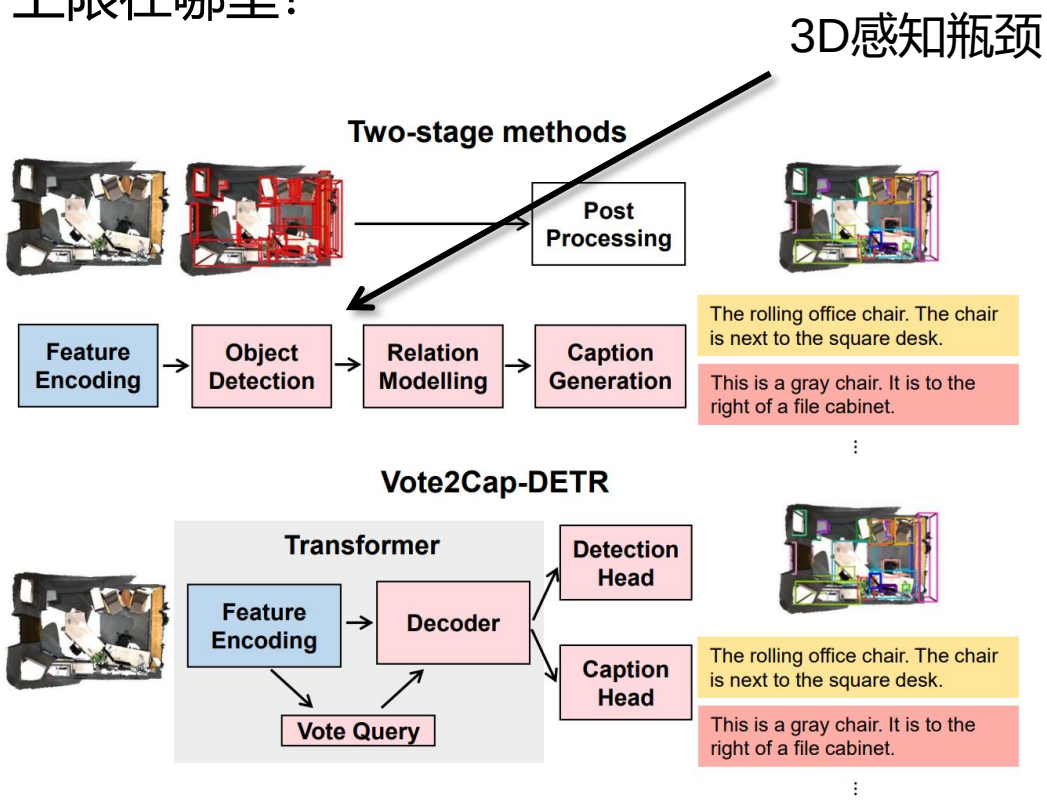
主要挑战：

1. 从稀疏的三维表征中预测准确的目标定位框；
2. 为场景中的物体生成以物体为中心的、富含信息量的目标描述文本；

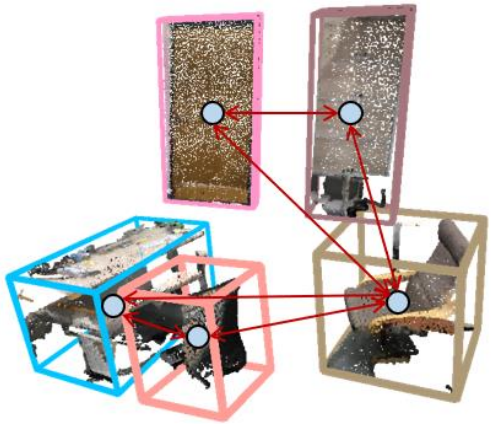
[1] Chen, Zhenyu, et al. "Scan2cap: Context-aware dense captioning in rgb-d scans." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

基于Transformer的端到端三维场景目标定位与描述生成

上限在哪里?



1. 3D检测器**误检**造成的累积误差
   i.e. miss / duplicated detection

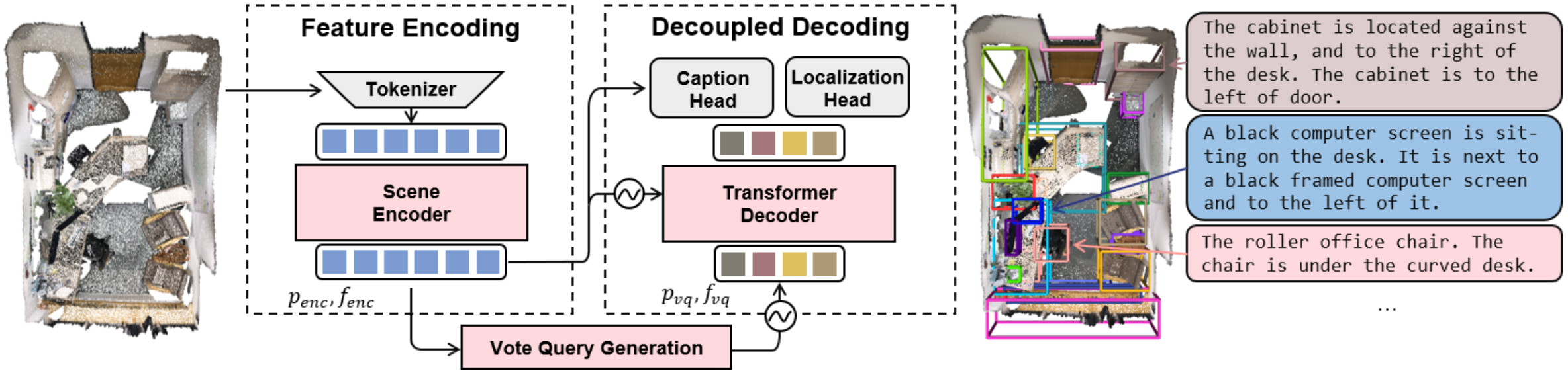


2. 场景**上下文信息**感知丢失
   i.e. room corners / textured walls

基于Transformer的端到端三维场景目标定位与描述生成

Vote2Cap-DETR，一个集合预测的视角：

1. 将一组点（点云）端到端翻译为一组“目标框——文本描述”对；
2. 使用解码器内的注意力，学习查询与查询间、以及查询与场景的交互；
3. 通过集合预测的训练手段，学习更强的物体表征。



基于Transformer的端到端三维场景目标定位与描述生成

如何将Transformer架构适配于三维场景？

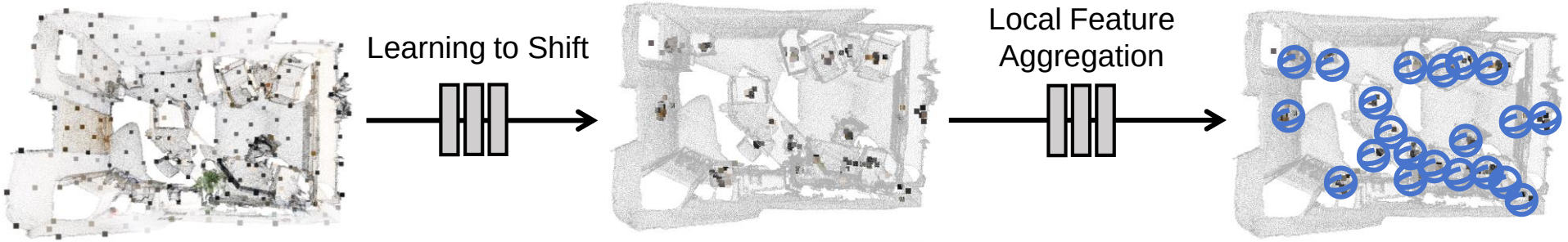
获得准确三维检测框有多难？

1. 自由度大 (x, y, z) vs 点云空间分布不均 (occupancy)；
2. 三维场景点云，物体表面采样稀疏 (sparsity)；

给目标查询 (object query) 多来一点先验！

空间先验
Spatial Prior

语义先验
Semantic Prior



最远点采样 p_{seed}

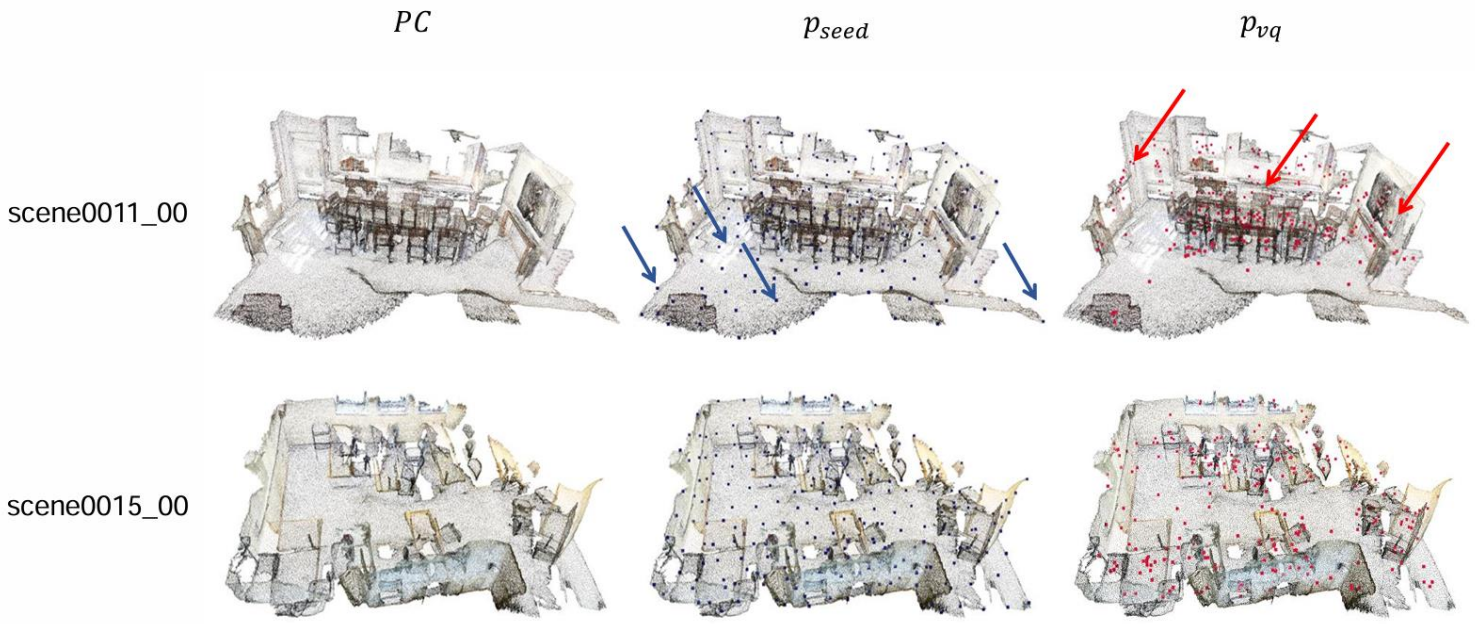
投票查询位置 p_{seed}

基于Transformer的端到端三维场景目标定位与描述生成

给目标查询 (object query) 多来一点先验!

空间先验
Spatial Prior

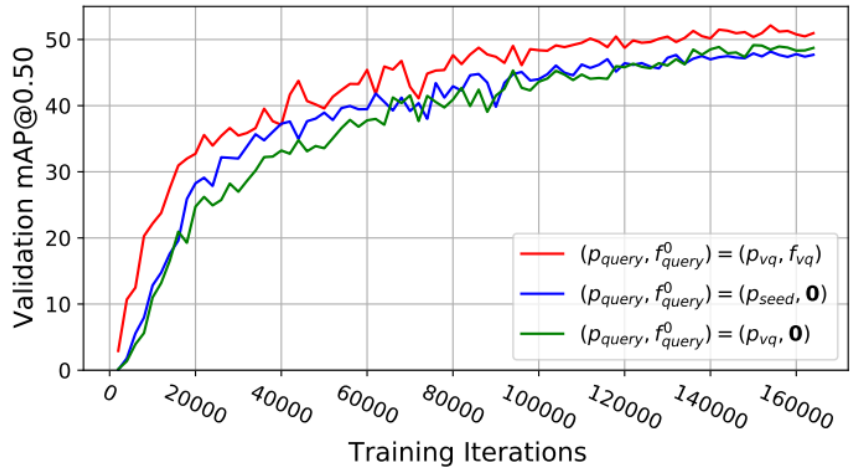
目标查询聚集于场景物体附近，带来了更强的感知物体能力



语义先验
Semantic Prior

大幅提高了第一层decoder的利用率

p_{query}	f^0_{query}	IoU=0.25		IoU=0.50		1st layer IoU=0.50	
		mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑
VoteNet Baseline		63.42	82.18	44.96	60.65	-	-
p_{seed}	**0**	67.25	84.91	48.18	64.98	34.80	55.06
p_{vq}	**0**	67.33	85.60	49.15	66.38	30.23	58.44
p_{vq}	f_{vq}	**69.61**	**87.20**	**52.13**	**69.12**	46.53	66.51

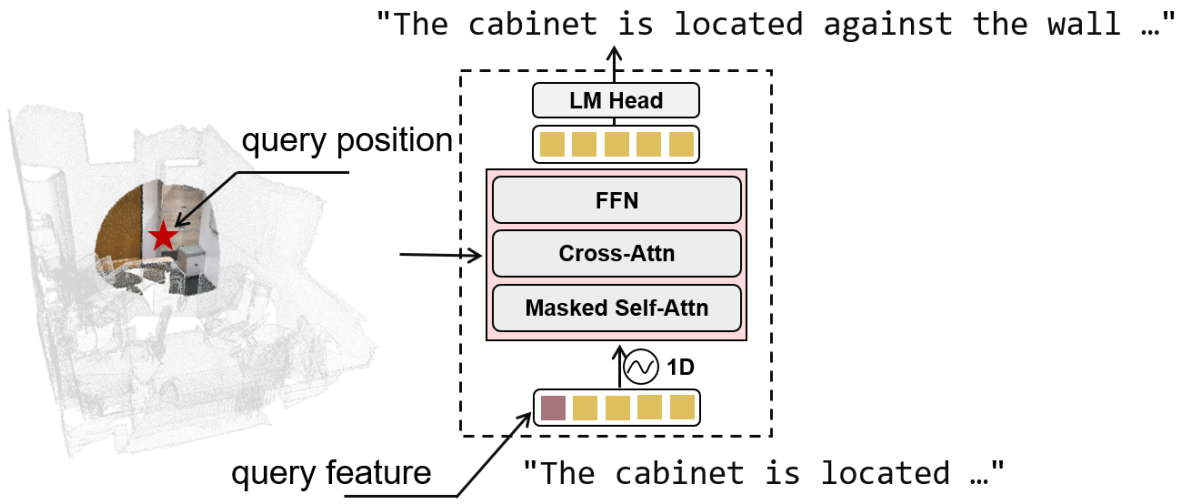


基于Transformer的端到端三维场景目标定位与描述生成

如何将Transformer架构适配于三维场景？

如何生成以物体为中心的、富含信息量的目标描述文本？

1. 怎样让模型聚焦于物体，并看到上下文特征 (local surroundings)
2. 如何区分同场景内的不同物体特征 (set-to-set training)

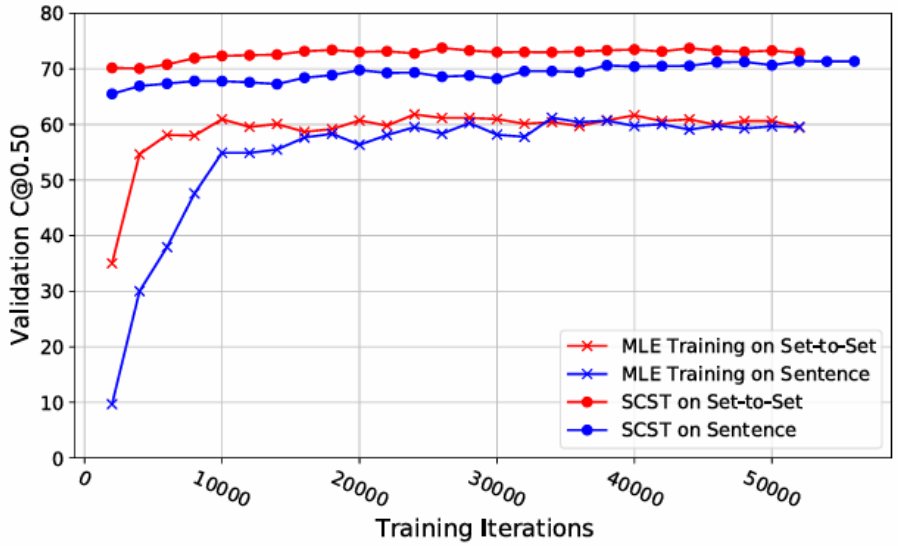


1. 用目标查询作为前缀，并引入局部上下文；
2. 一对一的匹配损失 (Set Loss)，学习更强的目标表征；

局部上下文的引入能够促使更准确的目标描述生成

key	IoU=0.25				IoU=0.5			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
-	68.62	38.61	27.67	58.47	60.15	34.02	25.80	53.82
global	70.05	39.23	27.84	58.44	61.20	34.66	25.93	53.79
local	**70.42**	**39.98**	**27.99**	**58.89**	**61.39**	**35.24**	**26.02**	**54.12**

集合预测的训练方式，促使更好的模型描述性能



基于Transformer的端到端三维场景目标定位与描述生成

评价指标：基于混合准确与查全的F1 score。

$$M^{\mathrm{F1}}@k\mathrm{IoU} = \frac{2 \times M^{\mathrm{P}}@k\mathrm{IoU} \times M^{\mathrm{R}}@k\mathrm{IoU}}{M^{\mathrm{P}}@k\mathrm{IoU} + M^{\mathrm{R}}@k\mathrm{IoU}}$$

Scan2Cap Benchmark

This table lists the benchmark results for the Scan2Cap Dense Captioning Benchmark scenario.

			Captioning F1-Score						Dense Captioning		Object Detection	
	Method	Info	CIDEr@0.5IoU		BLEU-4@0.5IoU		Rouge-L@0.5IoU		METEOR@0.5IoU		DCmAP		mAP@0.5	
→	vote2cap-detr	P	**0.3128** 1	+7.7%	**0.1778** 1	+3.6%	**0.2842** 1	+5.9%	**0.1316** 1	+2.8%	**0.1825** 1	+4.5%	**0.4454** 1	+2.6%
	Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Tao Chen, Gang YU, Taihao Li: End-to-End 3D Dense Captioning with Vote2Cap-DETR. CVPR 2023													
	CFM		0.2360 2		0.1417 2		0.2253 2		0.1034 2		0.1379 5		0.3008 5	
	CM3D-Trans+		0.2348 3		0.1383 3		0.2250 4		0.1030 3		0.1398 4		0.2966 7	
	Yufeng Zhong, Long Xu, Jiebo Luo, Lin Ma: Contextual Modeling for 3D Dense Captioning on Point Clouds.													
	Forest-xyz		0.2266 4		0.1363 4		0.2250 3		0.1027 4		0.1161 10		0.2825 10	
	D3Net - Speaker	P	0.2088 5		0.1335 6		0.2237 5		0.1022 5		0.1481 3		0.4198 2	
	Dave Zhenyu Chen, Qirui Wu, Matthias Niessner, Angel X. Chang: D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. 17th European Conference on Computer Vision (ECCV), 2022													
	3DJCG(Captioning)	P	0.1918 6		0.1350 5		0.2207 6		0.1013 6		0.1506 2		0.3867 3	
	Daigang Cai, Lichen Zhao, Jing Zhang†, Lu Sheng, Dong Xu: 3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds. CVPR2022 Oral													
	REMAN		0.1662 7		0.1070 7		0.1790 7		0.0815 7		0.1235 8		0.2927 9	
	NOAH		0.1382 8		0.0901 8		0.1598 8		0.0747 8		0.1359 6		0.2977 6	
	SpaCap3D	P	0.1359 9		0.0883 9		0.1591 9		0.0738 9		0.1182 9		0.3275 4	
	Heng Wang, Chaoyi Zhang, Jianhui Yu, Weidong Cai: Spatiality-guided Transformer for 3D Dense Captioning on Point Clouds. the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022													

scene0011_00

scene0015_00

3DJCG: This is a rectangular whiteboard. It is on the wall.

SpaCap3D: The whiteboard is affixed to the wall. It is to the right of the window.

Ours: The tv is on the wall. It is to the right of the table.

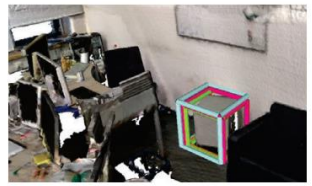
GT: This is a big black tv. It is above a thin table.

3DJCG: This is a brown table. It is in the middle of the room.

SpaCap3D: This is a wooden table. It is in the center of the room.

Ours: This is a wooden table. It is in the corner of the room.

GT: This is a small table with a wood look. It is the table closest to the front of the room in the upper left corner.



scene0025_00

scene0050_00

3DJCG: The is a small brown cabinet. It is to the right of the desk.

SpaCap3D: The cabinet is below the desk. It is to the left of the chair.

Ours: This is a white cabinet. It is to the right of the table.

GT: A white cabinet is sitting on the floor next to the wall. It is to the left of the couch.

3DJCG: This is a brown table. It is in front of the couch.

SpaCap3D: This is a wooden coffee table. It is in front of the couch.

Ours: This is a brown ottoman. It is to the right of the chair.

GT: This is a brown ottoman. It is in front of a couch.

IEEE TRANSACTIONS ON
PATTERN ANALYSIS AND
MACHINE INTELLIGENCE

基于任务解耦的端到端三维场景目标定位与描述生成

Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning, [T-PAMI 2024]

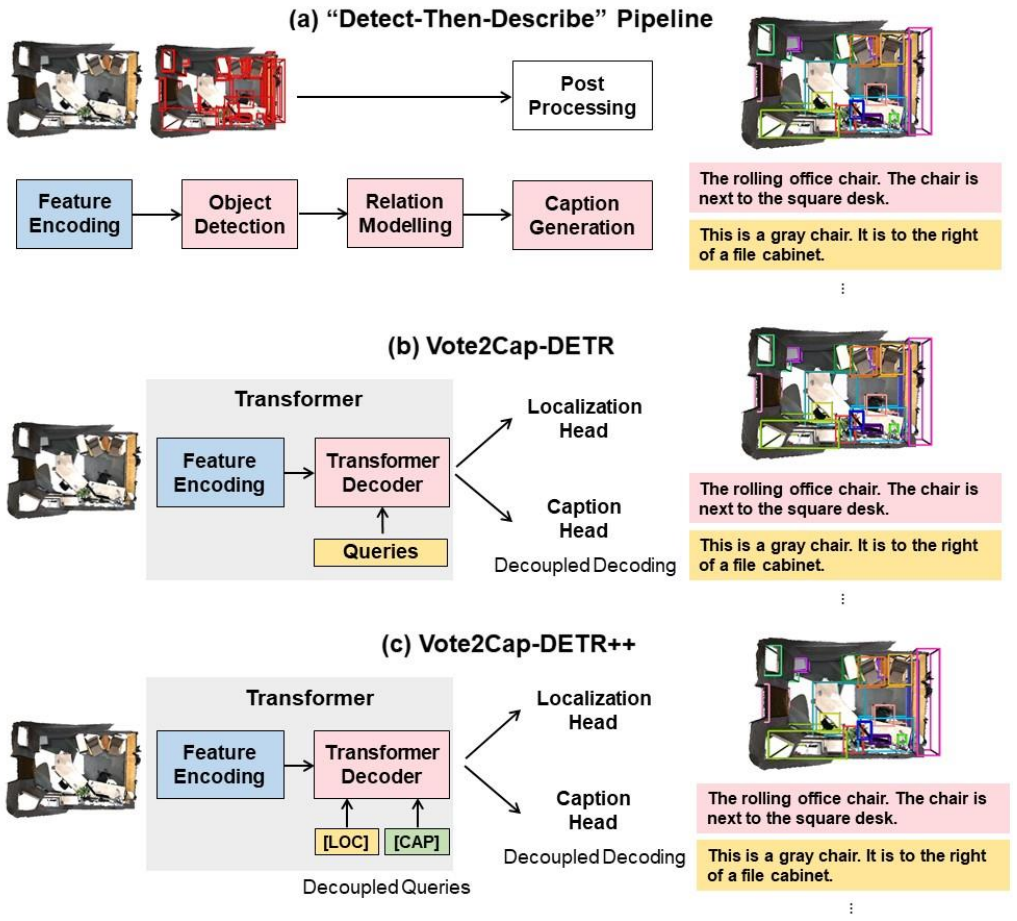
基于任务解耦的端到端三维场景目标定位与描述生成

对Vote2Cap-DETR结构的再思考：

1. 准确的目标定位，具备信息量的目标描述分别需要什么？
   - 目标定位：三维结构 (*3D structures*) 的感知；
   - 目标描述：高层次属性信息 (*attributes & relations*) 的感知；

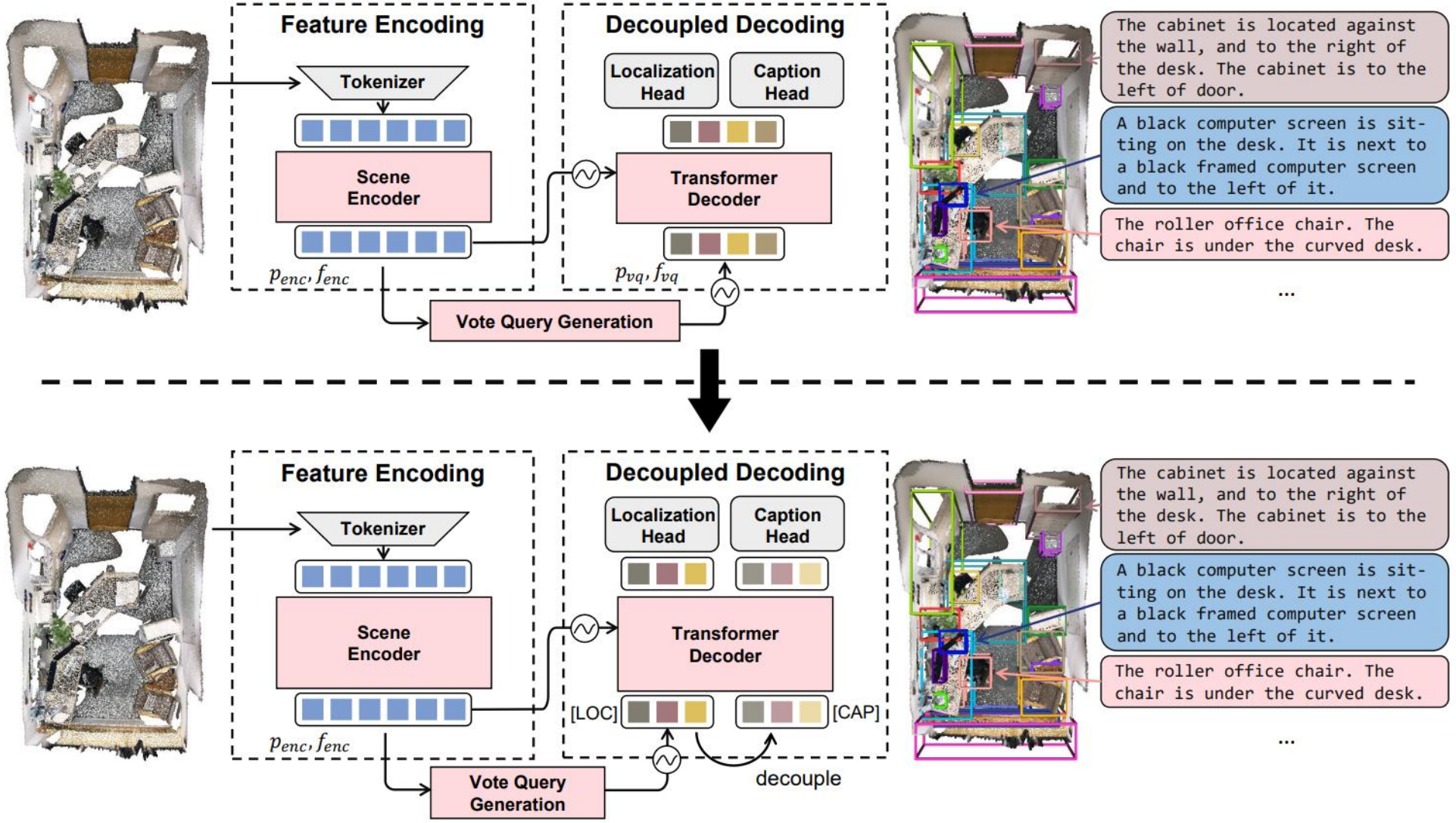
2. 现有 Transformer 架构的优势是什么？
   - 通过查询 (*querying*) 的形式获取目标特征；

3. 投票查询的空间先验是否太弱了？
   - 让空间偏置的学习更为简单！



基于任务解耦的端到端三维场景目标定位与描述生成

Vote2Cap-DETR (上图) vs. Vote2Cap-DETR++ (下图)

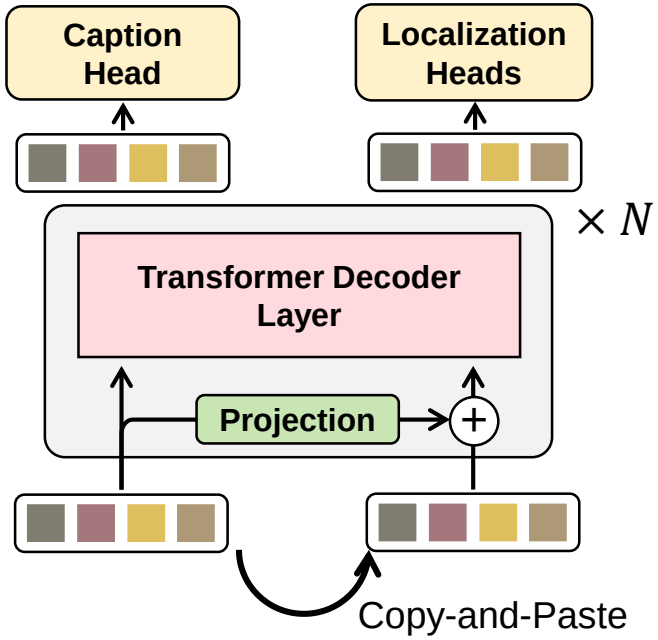


基于任务解耦的端到端三维场景目标定位与描述生成

任务解耦的目标查询

准确的目标定位，具备信息量的目标描述分别需要什么?

- 目标定位：三维结构 (*3D structures*) 的感知；
- 目标描述：高层次属性信息 (*attributes & relations*) 的感知；



1. 用复制黏贴的方式获得两组目标查询;
2. 不同查询用于不同任务;
3. 逐查询的特征映射 (query-wise projection);

1. 仅任务解耦：保证了检测性能，但文本描述准确性变差;
2. 查询对应：定位与描述任务都获得性能提升;

decouple	correspond	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	mAP@0.5↑
-	-	66.01	**37.41**	26.62	55.33	58.18
✓	-	65.38	37.07	26.71	55.41	58.67
✓	✓	**67.58**	37.05	**26.89**	**55.64**	**58.83**

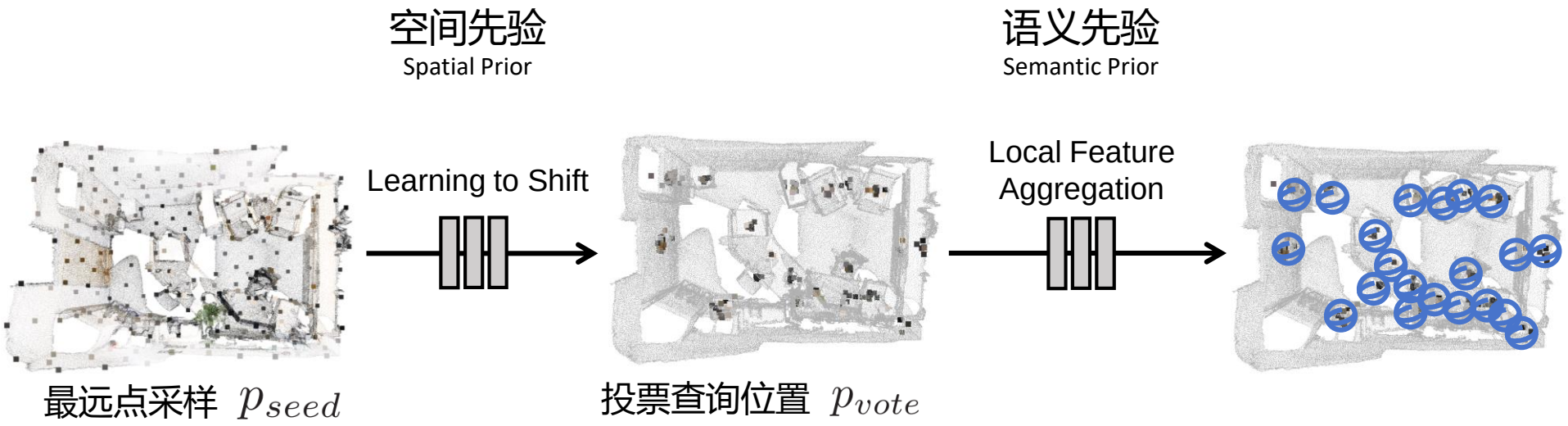
基于任务解耦的端到端三维场景目标定位与描述生成

回忆一下：投票查询

获得准确三维检测框有多难？

1. 自由度大 (x, y, z) vs 点云空间分布不均 (occupancy)；
2. 三维场景点云，物体表面采样稀疏 (sparsity)；

给目标查询 (object query) 多来一点先验！



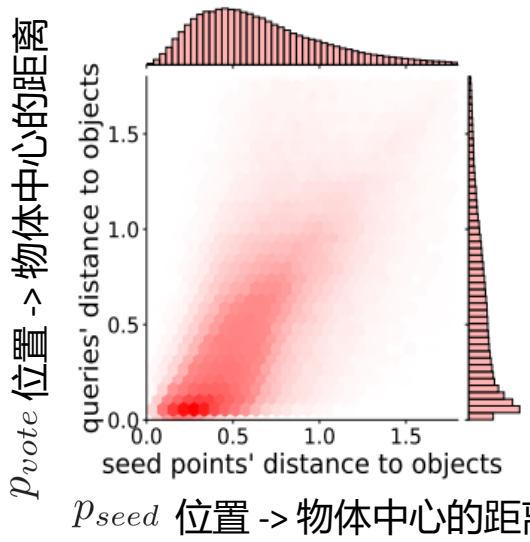
基于任务解耦的端到端三维场景目标定位与描述生成

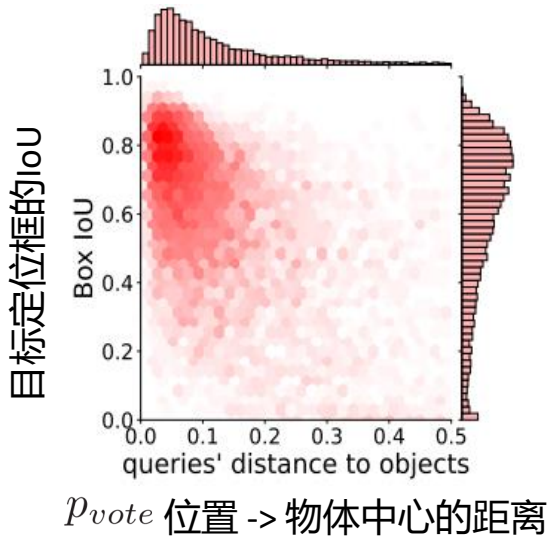
渐进式的投票优化策略

投票查询的空间先验是否太弱了？

- 让空间先验的学习更为简单！

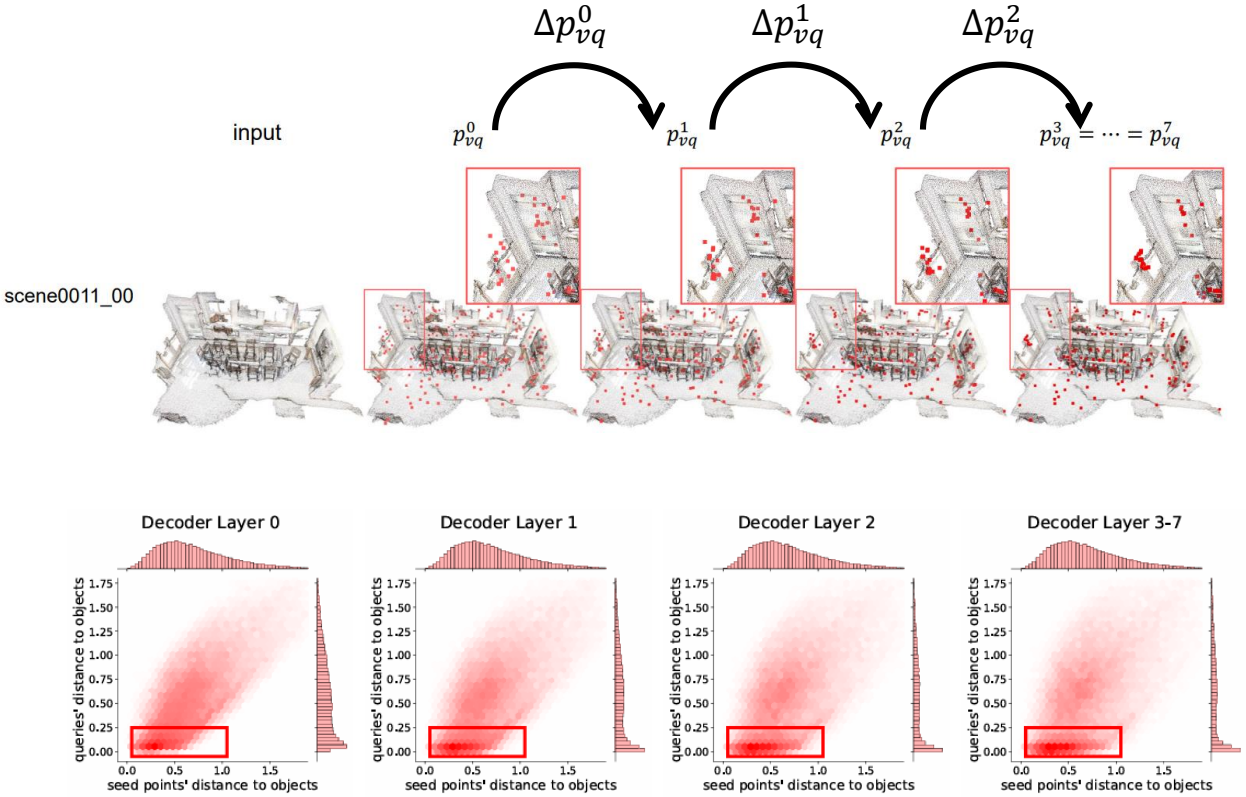
观察： 当目标查询离物体中心越近，它对应的目标框估计就越准





不如改变左图的斜率？

简单有效： 逐步去修正目标查询的空间位置



从第二层开始，有效获得性能提升

	Layer-id	0	1	2	3	4	5	6	7
mAP@0.5↑	Vote2Cap-DETR	48.17	49.91	51.20	52.11	52.50	52.26	52.31	52.49
	Vote2Cap-DETR++	48.07	51.57	53.66	54.34	55.11	55.34	55.52	55.48
	Δ	-0.10%	+1.66%	+2.46%	+2.23%	+2.61%	+3.12%	+3.21%	+2.99%

基于任务解耦的端到端三维场景目标定位与描述生成

评价指标：基于混合准确与查全的F1 score。

$$M^{\mathrm{F1}}@k\mathrm{IoU} = \frac{2 \times M^{\mathrm{P}}@k\mathrm{IoU} \times M^{\mathrm{R}}@k\mathrm{IoU}}{M^{\mathrm{P}}@k\mathrm{IoU} + M^{\mathrm{R}}@k\mathrm{IoU}}$$

Scan2Cap Benchmark

This table lists the benchmark results for the Scan2Cap Dense Captioning Benchmark scenario.

		Captioning F1-Score								Dense Captioning		Object Detection	
Method	Info	CIDEr@0.5IoU		BLEU-4@0.5IoU		Rouge-L@0.5IoU		METEOR@0.5IoU		DCmAP		mAP@0.5	
→ Vote2Cap-DETR++		**0.3360** 1	+2.3%	**0.1908** 1	+1.3%	**0.3012** 1	+1.7%	**0.1386** 1	+0.7%	**0.1864** 1	+0.4%	**0.5090** 1	+6.4%
Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, Tao Chen: Vote2Cap-DETR++: Decoupling Localization and Describing for End-to-End 3D Dense Captioning.													
vote2cap-detr	P	0.3128 2		0.1778 2		0.2842 3		0.1316 3		0.1825 2		0.4454 3	
Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Tao Chen, Gang YU, Taihao Li: End-to-End 3D Dense Captioning with Vote2Cap-DETR. CVPR 2023													
TMP		0.3029 3		0.1728 3		0.2898 2		0.1332 2		0.1801 3		0.4605 2	
CFM		0.2360 4		0.1417 4		0.2253 4		0.1034 4		0.1379 7		0.3008 7	
CM3D-Trans+		0.2348 5		0.1383 5		0.2250 6		0.1030 5		0.1398 6		0.2966 9	
Yufeng Zhong, Long Xu, Jiebo Luo, Lin Ma: Contextual Modeling for 3D Dense Captioning on Point Clouds.													
Forest-xyz		0.2266 6		0.1363 6		0.2250 5		0.1027 6		0.1161 12		0.2825 12	
D3Net - Speaker	P	0.2088 7		0.1335 8		0.2237 7		0.1022 7		0.1481 5		0.4198 4	
Dave Zhenyu Chen, Qirui Wu, Matthias Niessner, Angel X. Chang: D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. 17th European Conference on Computer Vision (ECCV), 2022													
3DJCG(Captioning)	P	0.1918 8		0.1350 7		0.2207 8		0.1013 8		0.1506 4		0.3867 5	
Daigang Cai, Lichen Zhao, Jing Zhang†, Lu Sheng, Dong Xu: 3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds. CVPR2022 Oral													

基于任务解耦的端到端三维场景目标定位与描述生成

ScanRefer

scene0011_00

3DJCG: "This is a rectangular whiteboard. It is on the wall."

SpaCap3D: "The whiteboard is affixed to the wall. It is to the right of the window."

Vote2Cap-DETR: "There is a rectangular black tv. It is on the wall over a table."

Vote2Cap-DETR++: "There is a rectangular black tv. It is on the wall to the right of the door."

Ground Truth: "There is a black square tv. Placed on the wall of the kitchen."

scene0025_00

3DJCG: "The is a small brown cabinet. It is to the right of the desk."

SpaCap3D: "The cabinet is below the desk. It is to the left of the chair."

Vote2Cap-DETR: "This is a white cabinet. It is to the right of a chair."

Vote2Cap-DETR++: "This is a white cabinet. It is to the right of a chair."

Ground Truth: "The file cabinet is located to the left of the couch. The file cabinet is also located to the right of the chair."

scene0015_00

3DJCG: "This is a brown table. It is in the middle of the room."

SpaCap3D: "This is a wooden table. It is in the center of the room."

Vote2Cap-DETR: "This is a wooden table. It is in the corner of the room."

Vote2Cap-DETR++: "The table is in the corner of the room. It is to the right of the chair."

Ground Truth: "This is a small table with a wood look. It is the table closest to the front of the room in the upper left corner."

scene0246_00

3DJCG: "The curtain is on the left side of the room. It is to the left of the curtain."

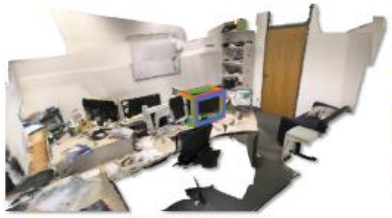
SpaCap3D: "The curtain is white. It is in the far right corner of the room."

Vote2Cap-DETR: "There is a rectangular curtain. It is to the left of the bed."

Vote2Cap-DETR++: "The curtain is on the wall. It is to the left of the table."

Ground Truth: "This is a red curtain. It is to the left of a table."

Nr3D



scene0025_00

SpaCap3D: "The keyboard closest to the door."

Vote2Cap-DETR: "The monitor closest to the door."

Vote2Cap-DETR++: "The monitor closest to the door."

Ground Truth: "Its the closest monitor to the door."

scene0169_00

SpaCap3D: "The long table with two chairs at it."

Vote2Cap-DETR: "The larger of the two tables."

Vote2Cap-DETR++: "The correct cabinet is in the corner of the room."

Ground Truth: "it is the large cabinet on the wall under the painting."

scene0046_00

SpaCap3D: "The backpack closest to the door."

Vote2Cap-DETR: "The backpack on the floor next to the bed."

Vote2Cap-DETR++: "The backpack closest to the window."

Ground Truth: "The backpack that is between the bed and window"

scene0300_00

SpaCap3D: "The bag on the table."

Vote2Cap-DETR: "The monitor closest to the door."

Vote2Cap-DETR++: "The monitor closest to the window."

Ground Truth: "The monitor that is next to the windows"

基于大语言模型的交互式三维视觉语言通才模型

LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning, [CVPR 2024]

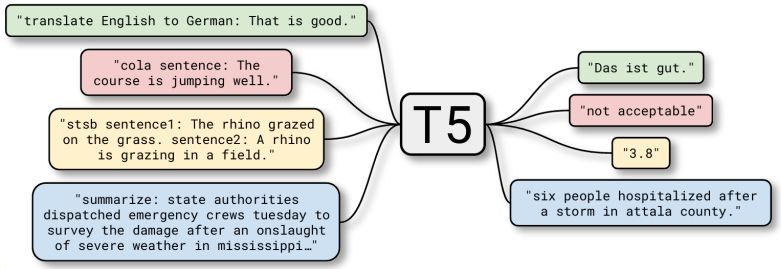
基于大语言模型的交互式三维视觉语言通才模型

大语言通才模型

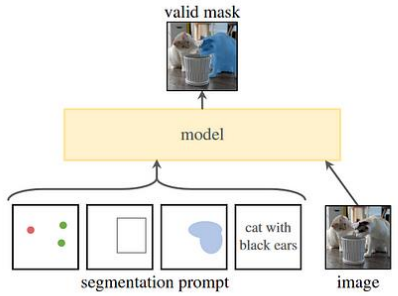
OpenAI ChatGPT4.0

LLAMA 2

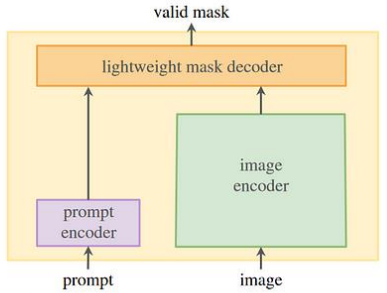
Qwen



可交互的视觉感知

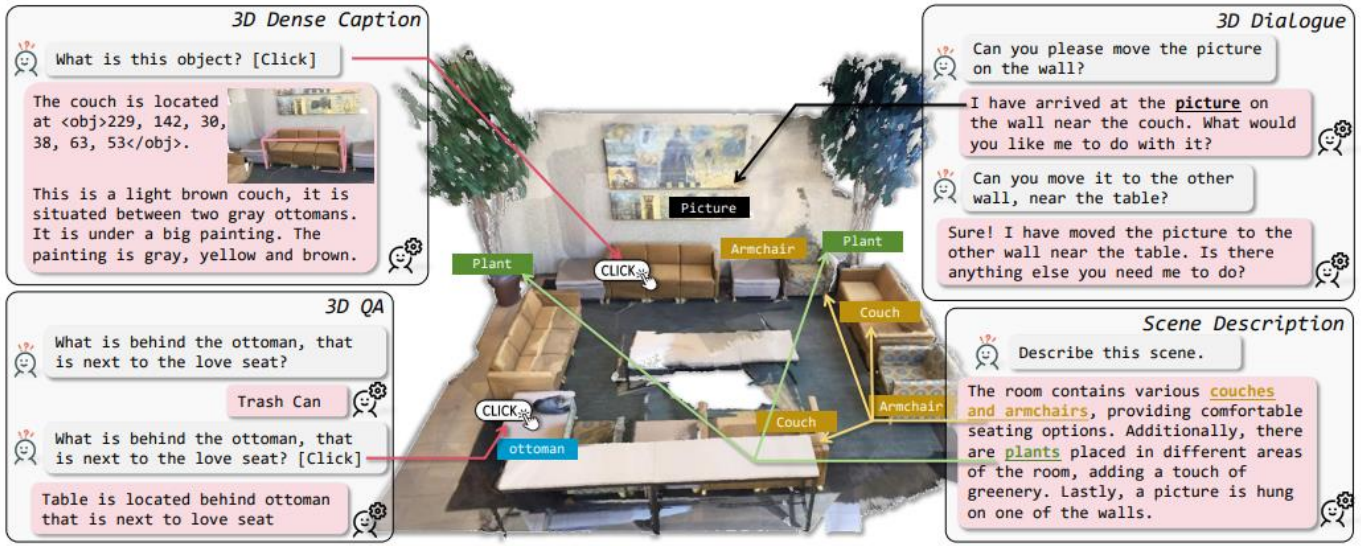


(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)

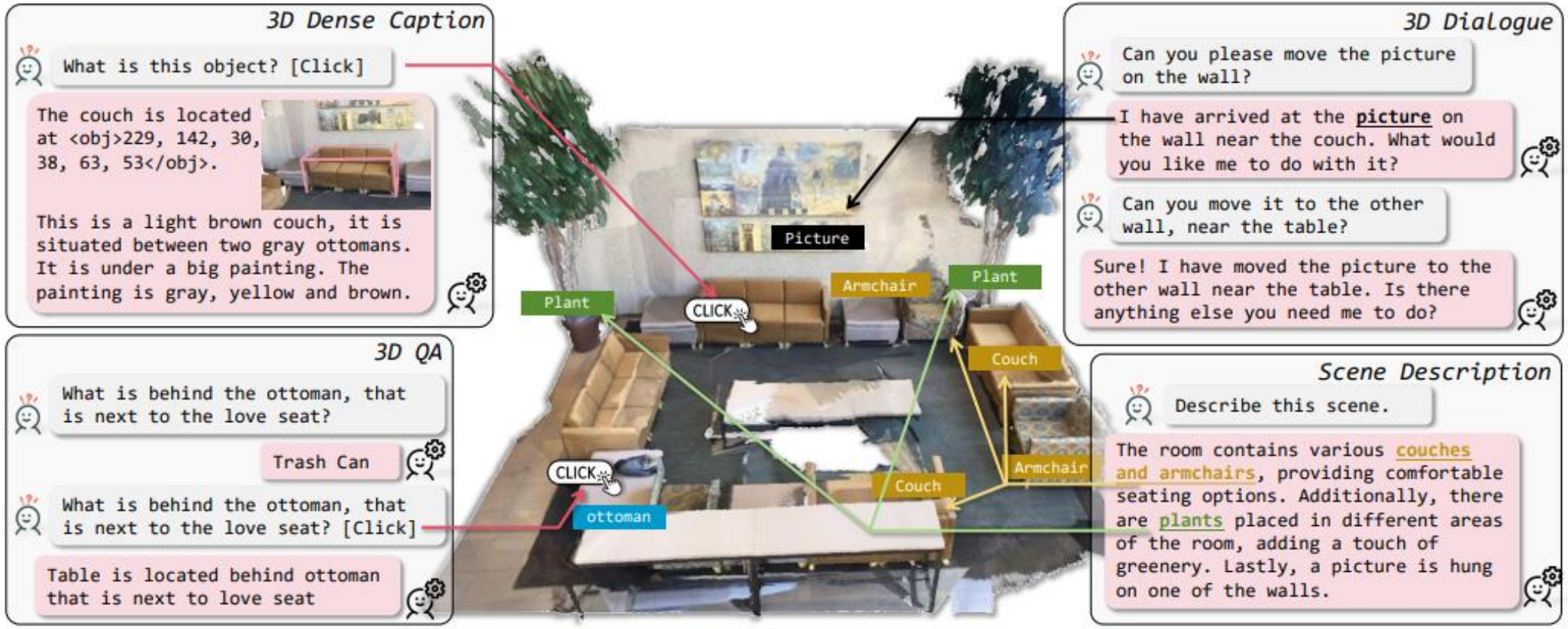
在三维场景下的可交互感知、推理、与规划



基于大语言模型的交互式三维视觉语言通才模型

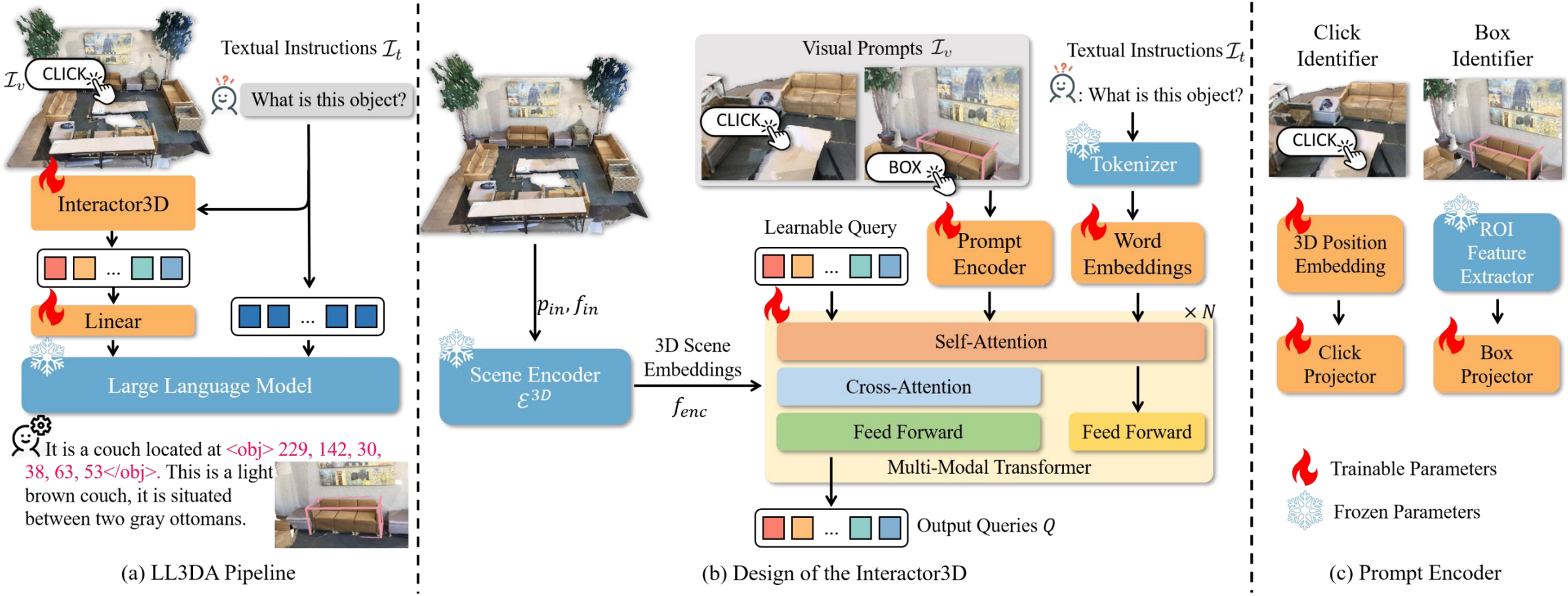
主要挑战：

1. 让大语言模型理解三维场景信息；
2. 三维点云的排序不变性 *vs.* 因果语言模型 (Causal LM) 的时序；
3. 三维场景理解任务粒度跨度大，需要潜在的视觉交互：场景内物体、完整场景；



基于大语言模型的交互式三维视觉语言通才模型

1. 为大语言模型学习统一的、交互可知的三维场景表征；
2. 平衡单模态专家模型间的模态差异；
3. 将排序不变的三维特征序列化，与大语言模型特征空间对齐；



基于大语言模型的交互式三维视觉语言通才模型

LL3DA展现出强大的语言问答、目标描述能力

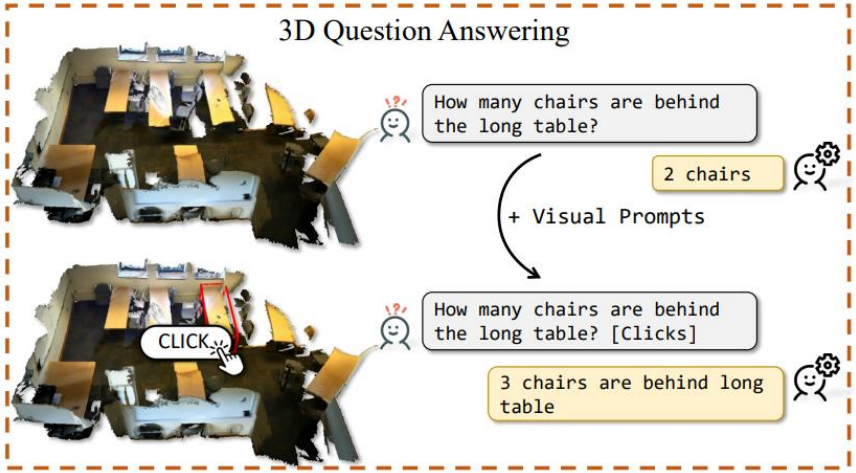
Table 1. State-of-the-Art 3D Dense Captioning.

Method	ScanRefer								Nr3D			
	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap[11]	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[29]	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42	-	-	-	-
SpaCap3D[51]	-	-	-	-	44.02	25.26	22.33	45.36	33.71	19.92	22.61	50.50
REMAN[38]	62.01	36.37	26.76	56.25	45.00	26.31	22.67	46.96	34.81	20.37	23.01	50.99
D3Net[7]	-	-	-	-	46.07	30.29	24.35	51.67	33.85	20.70	23.13	53.38
Contextual[62]	-	-	-	-	46.11	25.47	22.64	45.96	35.26	20.42	22.77	50.78
UniT3D[12]	-	-	-	-	46.69	27.22	21.91	45.98	-	-	-	-
3DJCG[4]	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80	38.06	22.82	23.77	52.99
3D-VLP[30]	70.73	41.03	28.14	**59.72**	54.94	32.31	24.83	51.51	-	-	-	-
3D-VisTA*[65]	-	-	-	-	61.60	34.10	**26.80**	55.00	-	-	-	-
Vote2Cap-DETR[9]	71.45	39.34	**28.25**	59.33	61.81	34.46	26.22	54.40	43.84	26.68	25.41	54.43
LL3DA (Ours)	**74.17**	**41.41**	27.76	59.53	**65.19**	**36.79**	25.97	**55.06**	**51.18**	**28.75**	**25.91**	**56.61**

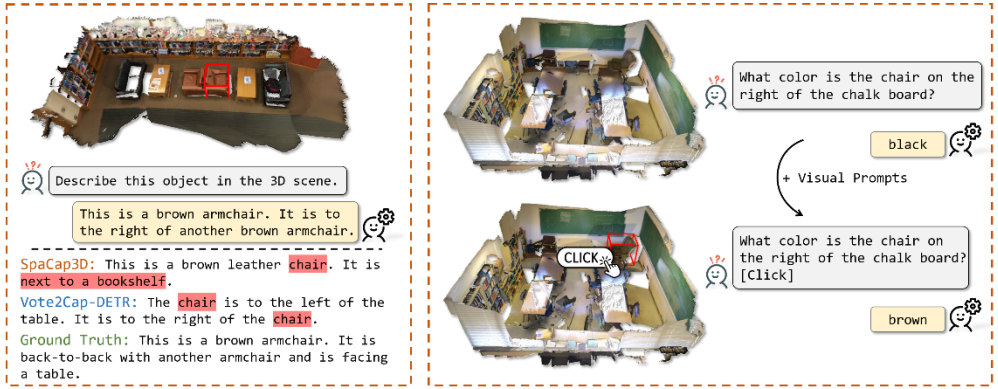
Table 2. State-of-the-Art 3D Question Answering. (**w/o** visual prompts during inference)

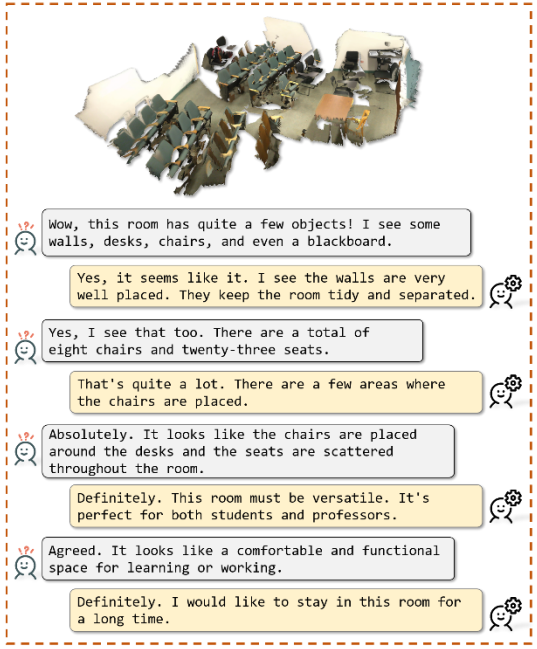
Method	Answer Type	Validation				Test w/ object				Test w/o object			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
ScanQA[2]	CLS	64.86	10.08	13.14	33.33	67.29	12.04	13.55	34.34	60.24	10.75	12.59	31.09
Clip-Guided[43]		-	-	-	-	69.53	**14.64**	13.94	35.15	62.83	11.73	13.28	32.41
Multi-CLIP[17]		-	-	-	-	68.70	12.65	13.97	35.46	63.20	12.87	13.36	32.61
3D-VLP[30]		66.97	11.15	13.53	34.51	70.18	11.23	14.16	35.97	63.40	**15.84**	13.13	31.79
3D-VisTA[65]		-	-	-	-	68.60	10.50	13.80	35.50	55.70	8.70	11.69	29.60
3D-LLM*[26]	GEN	69.40	12.00	14.50	35.70	69.60	11.60	14.90	35.30	-	-	-	-
LL3DA (Ours)		**76.79**	**13.53**	**15.88**	**37.31**	**78.16**	13.97	**16.38**	**38.15**	**70.29**	12.19	**14.85**	**35.17**

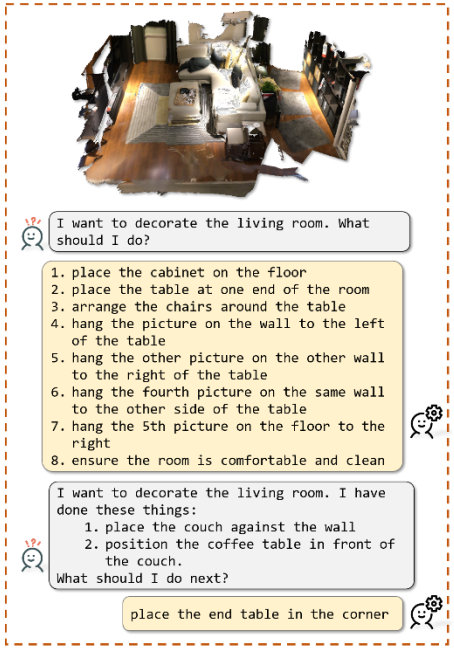
使用视觉交互，去除自然文本查询中的歧义

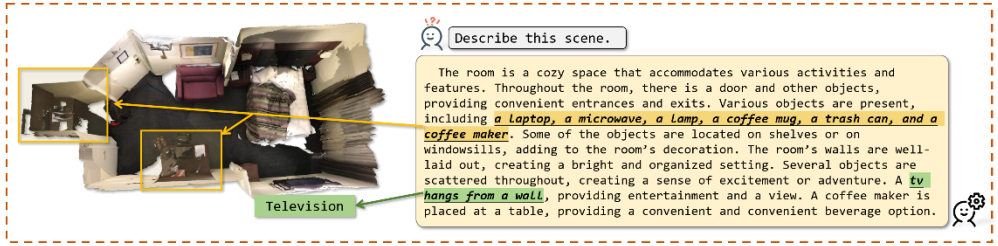


基于大语言模型的交互式三维视觉语言通才模型









除了高质量文本输出外：

给定视觉提示，以文本形式输出坐标和类别，实现三维开放词汇检测

What is this object? [Click]

the door is localized at <obj>241, 72, 101, 13, 39, 184</obj>

Method	ScanNet Vocabulary				ScanNet200 Vocabulary			
	IoU=0.25		IoU=0.5		IoU=0.25		IoU=0.5	
	mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑
VoteNet [45]	57.17	81.18	31.50	50.08	-	-	-	-
Ours	48.94	65.15	32.48	49.46	7.40	12.10	5.20	9.04

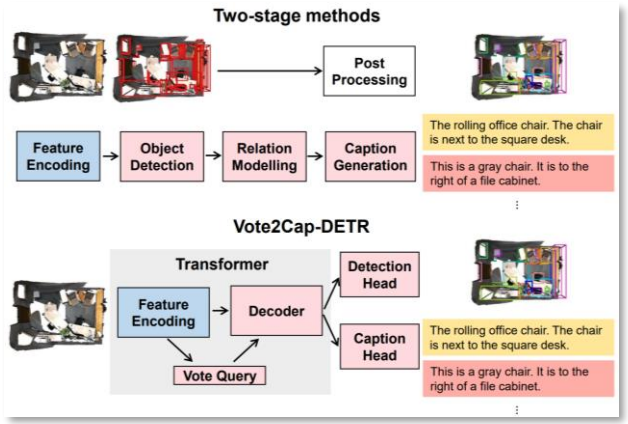
Input 3D Scene | VoteNet | LL3DA | Ground Truth Annotation

with some visual prompts…

总结与展望

Conclusions and Future Work

总结与展望——总结

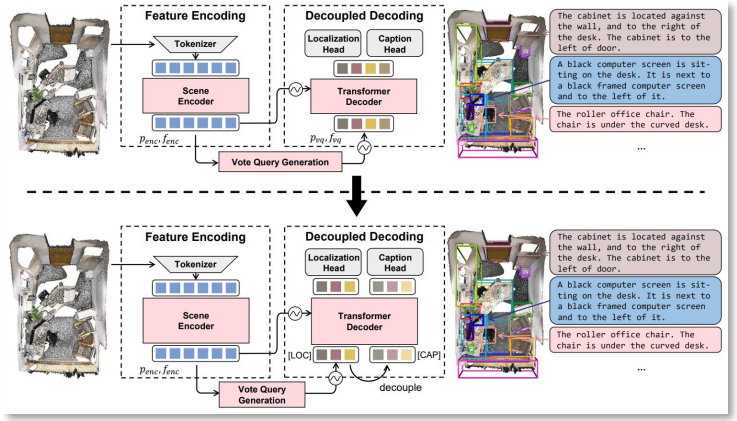


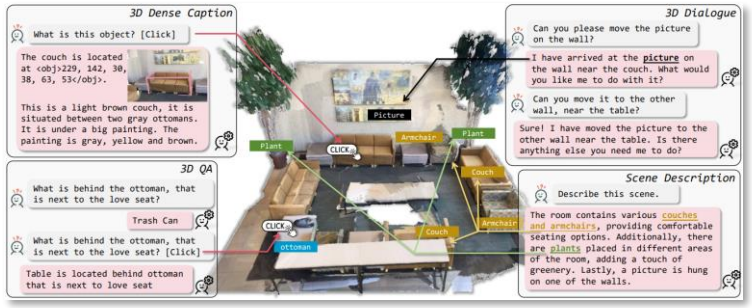
端到端的三维场景目标定位与描述生成：

1. 打破“检测—再—描述”框架的三维场景感知瓶颈；
2. 引入空间与语义先验的投票查询；
3. 集合到集合的训练策略；

目标定位与描述生成的任务解耦：

1. 解耦的目标特征查询；
2. 渐进式的目标查询空间偏置修正；





可交互的三维大语言通才模型：

1. 统一自回归架构的多任务通才模型；
2. 在语言指令的基础上，处理视觉交互；
3. 基于视觉交互，使用语言输出准确坐标；

总结与展望——展望

What's Next：

1. 文本生成的**幻觉**问题 (hallucination)——生成毫无逻辑的回答

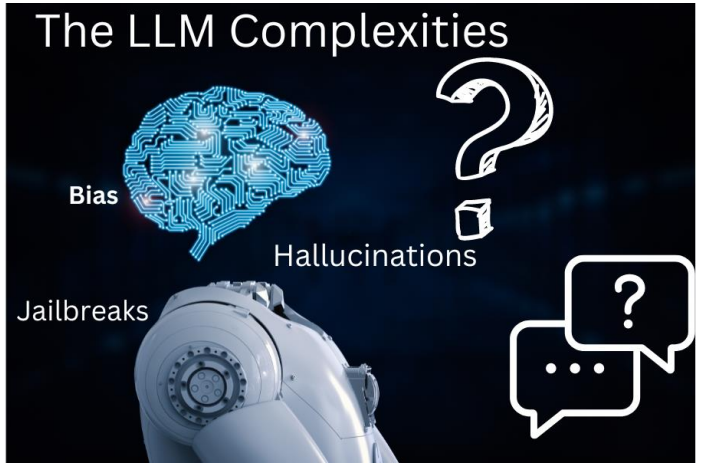
   Next-token prediction 以外的其他监督方式：RLHF？

2. 点云——有限的**三维表征**——损失小物体的感知

   稠密三维表征？3D Gaussian Splatting？混合表征？

3. 可**动态交互**的三维场景理解——只能基于固定场景，难以落实实际应用

   第一人称数据？Point cloud video？多传感器时序数据？Embodied Agents？

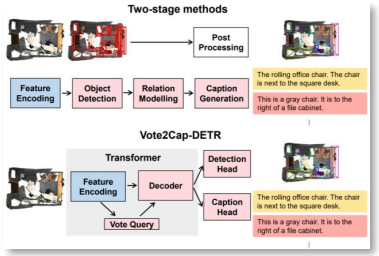


硕士期间的主要研究成果

Publications, Achievements, and Invited Talks

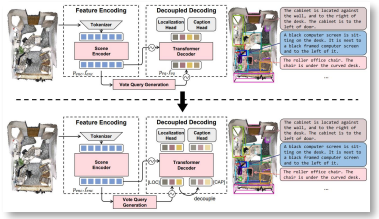
硕士期间的主要研究成果

Language for Better 3D Perception



[CVPR 2023, 1st author] **Vote2Cap-DETR**

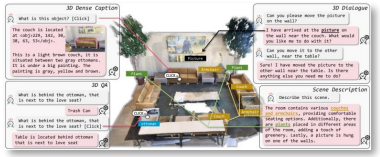
- **TLNR:** An set-to-set perspective towards localizing and describing objects from 3D scenes.
- 1st place of the Scan2Cap Challenge @ ICCV 2023.



[T-PAMI 2024, 1st author] **Vote2Cap-DETR++**

- **TLNR**: Decoupled feature extraction for better object localizing and describing.

Large Language 3D Assistants



[CVPR 2024, 1st author] **LL3DA**

- **TLNR:** A large language 3D assistant responding to text and visual interactions in 3D scenes.

[under review] **M3DBench** (4th Author)

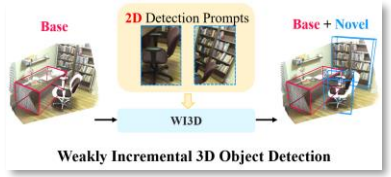
- **TLNR:** A benchmark querying 3D LLMs with diverse tasks and visual prompts.

Generative 3D Foundation Models

[under review, 1st author] **MeshXL**

- **TLNR:** Generative pre-trained 3D foundation models for auto-regressive 3D mesh generation.

Class-Incremental 3D Detection



[under review] **WI3D** (2nd Author)

- **TLNR:** Learn to detect new categories with 2D images.

Invited Talks

VALSE 2023 无锡，中国

“End-to-End 3D Dense Captioning with Vote2Cap-DETR”

ICCV 2023 巴黎，法国

“Vote2Cap-DETR: A Set-to-Set Perspective Towards 3D Dense Captioning”

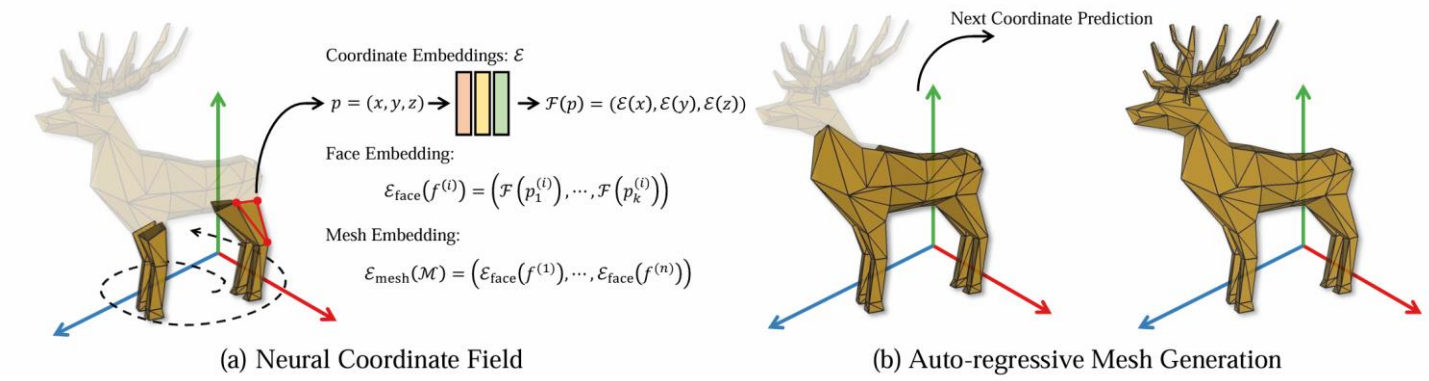
Reviewer services: NeurIPS 2024

To be continued …

一点小私货

Paper Link:

[TL;NR] Sequence modelling for 3D Mesh. Building generative pre-trained 3D Foundation Models for Auto-regressive Mesh Generation.



(a) Neural Coordinate Field (b) Auto-regressive Mesh Generation

MeshXL: Neural Coordinate Field for Generative 3D Foundation Models

Sijin Chen[1,2,*] Xin Chen[1,†] Anqi Pang[1] Xianfang Zeng[1] Yijun Fu[1]
Wei Cheng[1] Fukun Yin[1,2] Yanru Wang[1] Zhibin Wang[1]
Jingyi Yu[3] Gang Yu[1] Bin Fu[1] Tao Chen[2,‡]

[1]Tencent PCG [2]Fudan University [3]ShanghaiTech University
† project lead ‡ corresponding author

Figure 1: **We present MeshXL, a family of generative pre-trained transformers, for the direct generation of 3D meshes.** We validate that Neural Coordinate Field, an explicit coordinate representation with implicit neural embeddings, is a simple-yet-effective sequence representation for large-scale mesh modelling.

Abstract

The polygon mesh representation of 3D data exhibits great flexibility, fast rendering speed, and storage efficiency, which is widely preferred in various applications. However, given its unstructured graph representation, the direct generation of high-fidelity 3D meshes is challenging. Fortunately, with a pre-defined ordering strategy, 3D meshes can be represented as sequences, and the generation process can be seamlessly treated as an auto-regressive problem. In this paper, we validate the **Neur**al **C**oordinate **F**ield (NeurCF), an explicit coordinate representation with implicit neural embeddings, is a simple-yet-effective representation for large-scale sequential mesh modelling. After that, we present MeshXL, a family of generative pre-trained auto-regressive models, which addresses the process of 3D mesh generation with modern large language model approaches. Extensive

*Research done when Sijin Chen was a Research Intern at Tencent PCG.

Preprint. Under review.

感谢陪伴

Supervisor

Tao Chen

Collaborators

Hongyuan Zhu

Gang Yu

Xin Chen

Chi Zhang

Mingsheng Li

Peng Guo

Chuangguan Ye