GroundFlow: A Plug-in Module for Temporal Reasoning on 3D Point Cloud Sequential Grounding

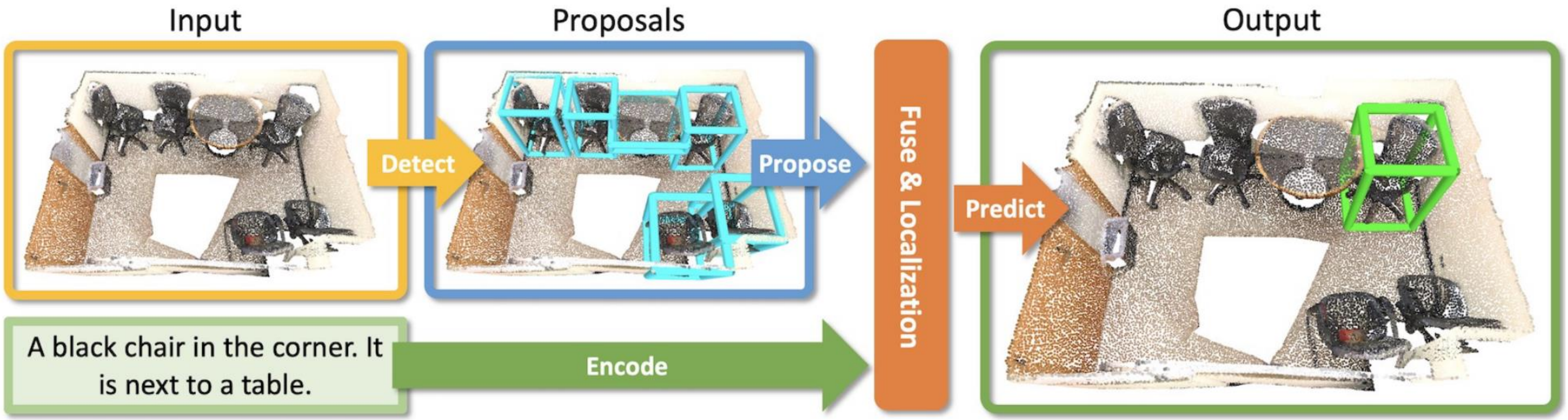
Zijun Lin[1,3] Shuting He[2] Cheston Tan[3] Bihan Wen[1]

[1]Nanyang Technological University [2]Shanghai University of Finance and Economics
[3]Centre for Frontier AI Research, A*STAR

ICCV HONOLULU HAWAII
OCT 19-23, 2025

Background – 3D Visual Grounding Dataset (3DVG)

ScanRefer



Multi3DRefer

ScanRefer: 3D Object Localization in rgb-d scans using natural language. ECCV, 2020.

Multi3drefer: Grounding text description to multiple 3d objects. ICCV, 2023.

Task-Oriented Sequential Grounding in 3D scenes (SG3D)

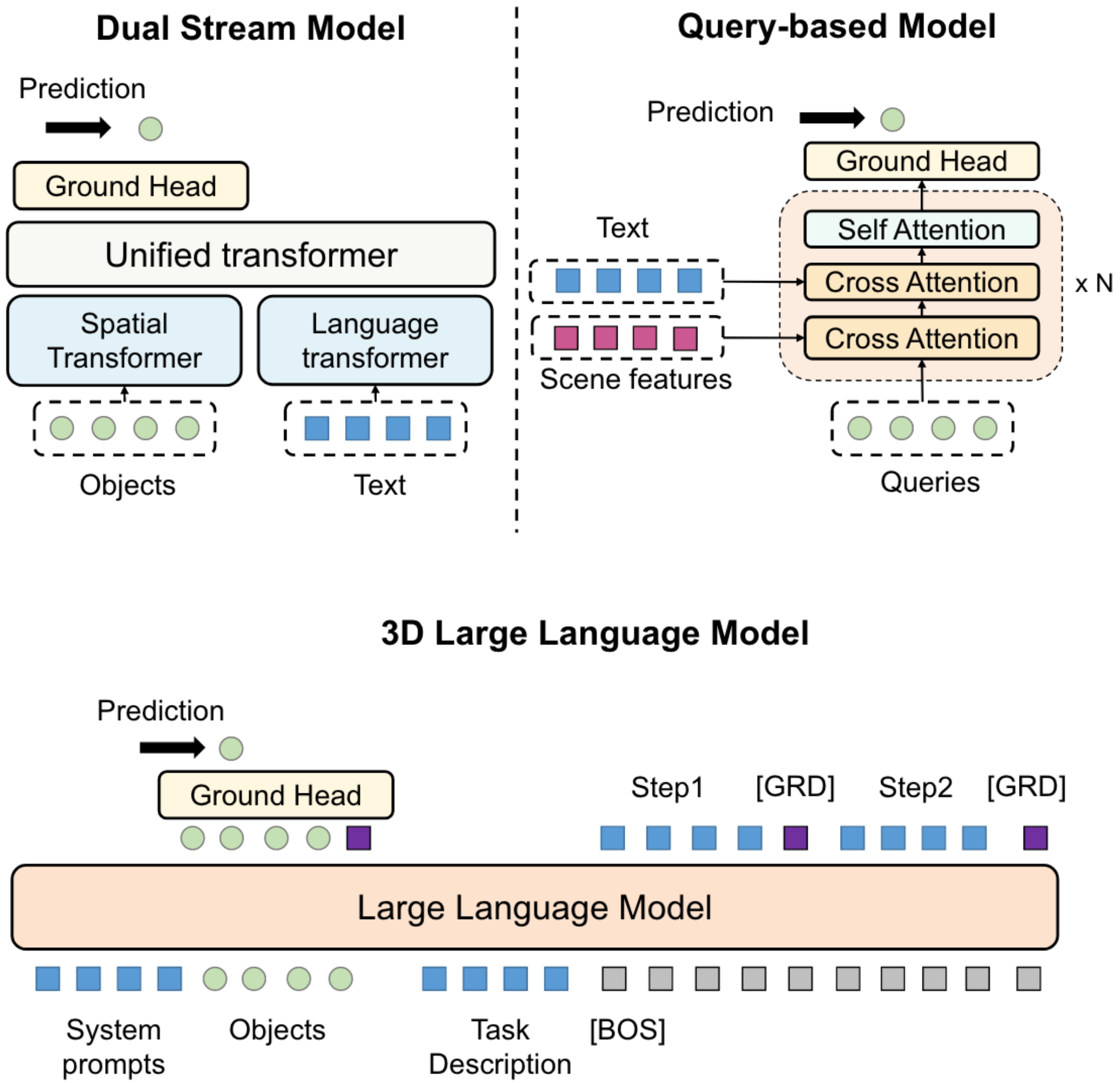
Task-Oriented: Unlike traditional object-centric visual grounding, the object is not explicitly described in the SG3D instruction.

Sequential Grounding: Require the agent to understand the context, remember the history steps and output sequence of BBox with order.

Use pronouns (‘it’, ‘them’, ‘here’, and ‘the other’, etc.) as much as possible to make the step concise

Task-oriented Sequential Grounding and Navigation in 3D Scenes. Arxiv:2408.04034

Current Sequential Grounding Methods



Task-oriented Sequential Grounding and Navigation in 3D Scenes. Arxiv:2408.04034

Current Sequential Grounding Methods

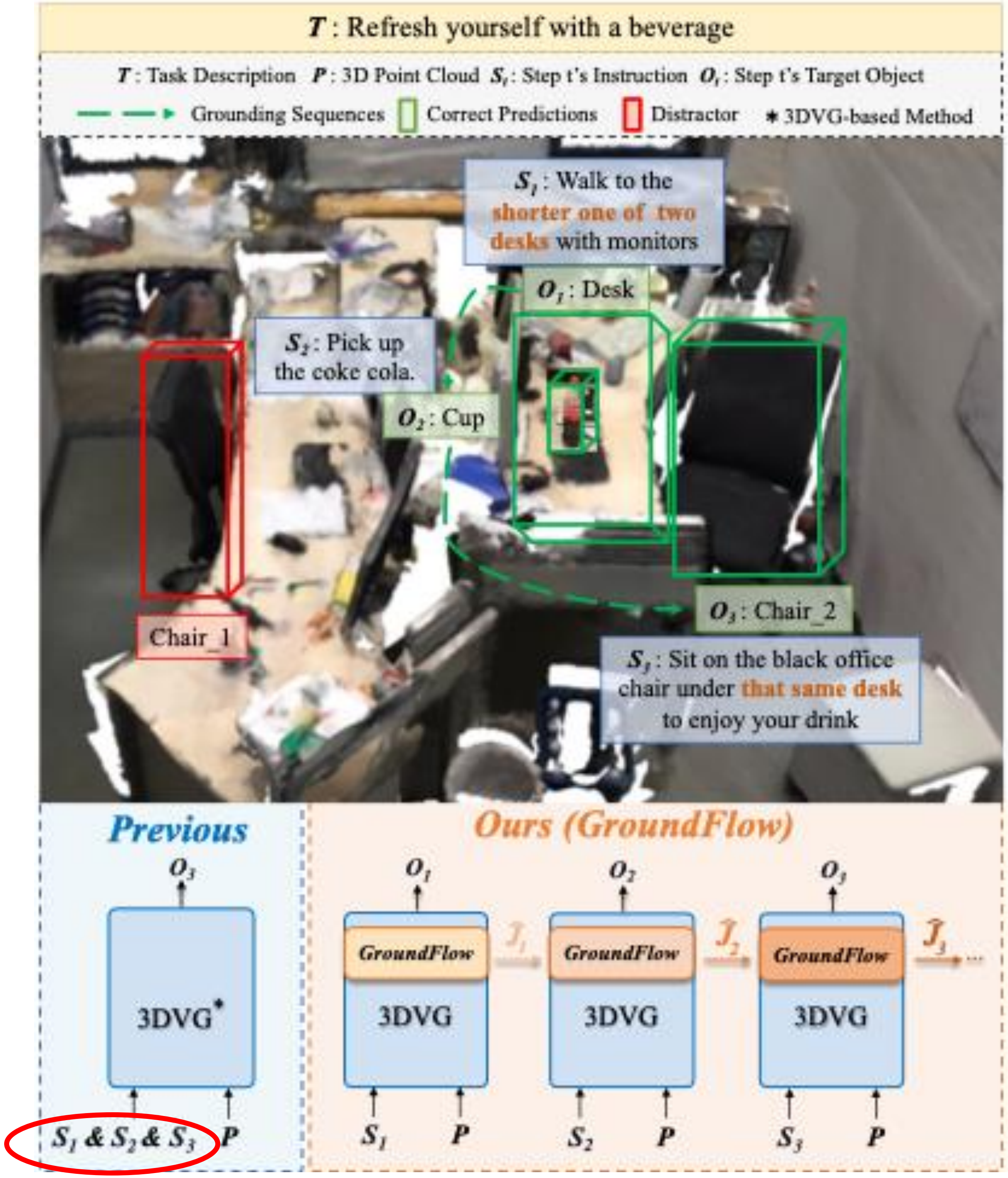
P: 3D Point Cloud

S_t: Step t's Text Instruction

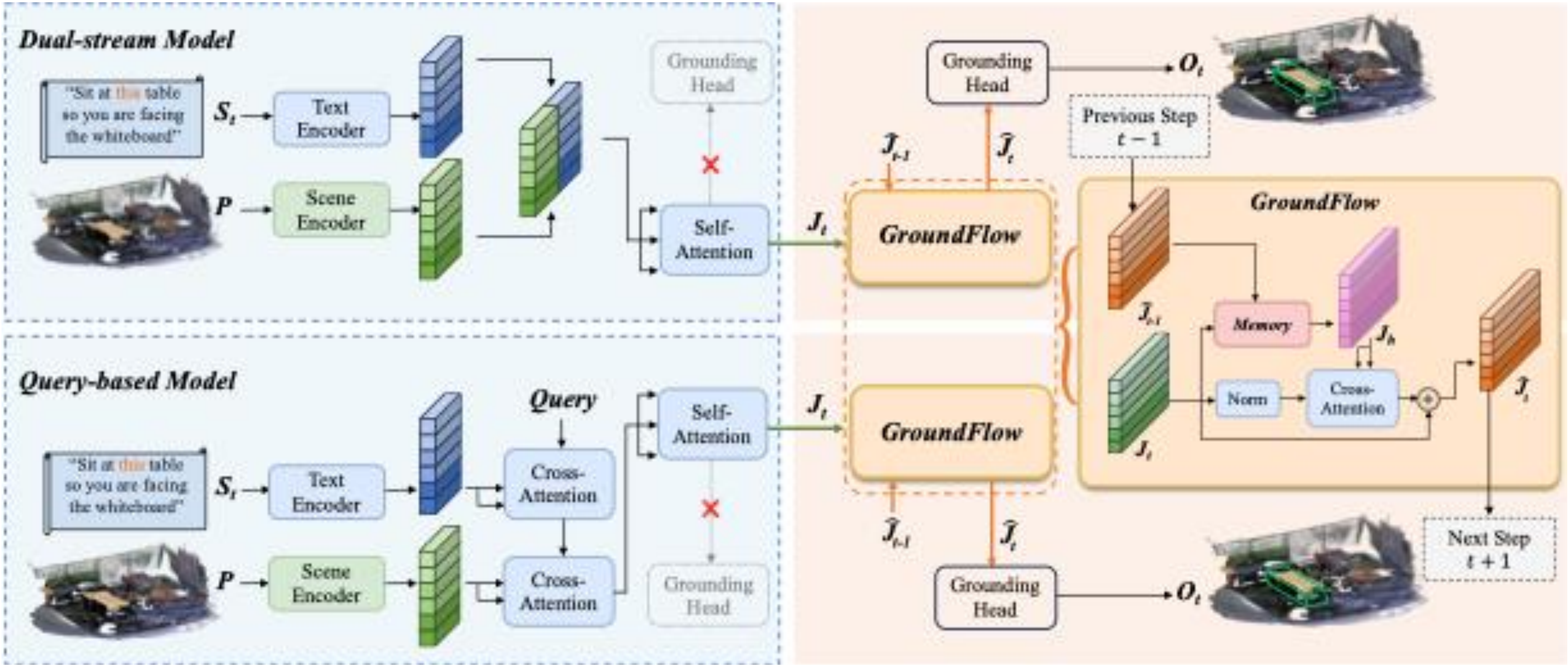
J_t: Step t's Joint Embedding

O_t: Step t's Target Object

Not designed to reason over historical information, important past information mixes with irrelevant details →



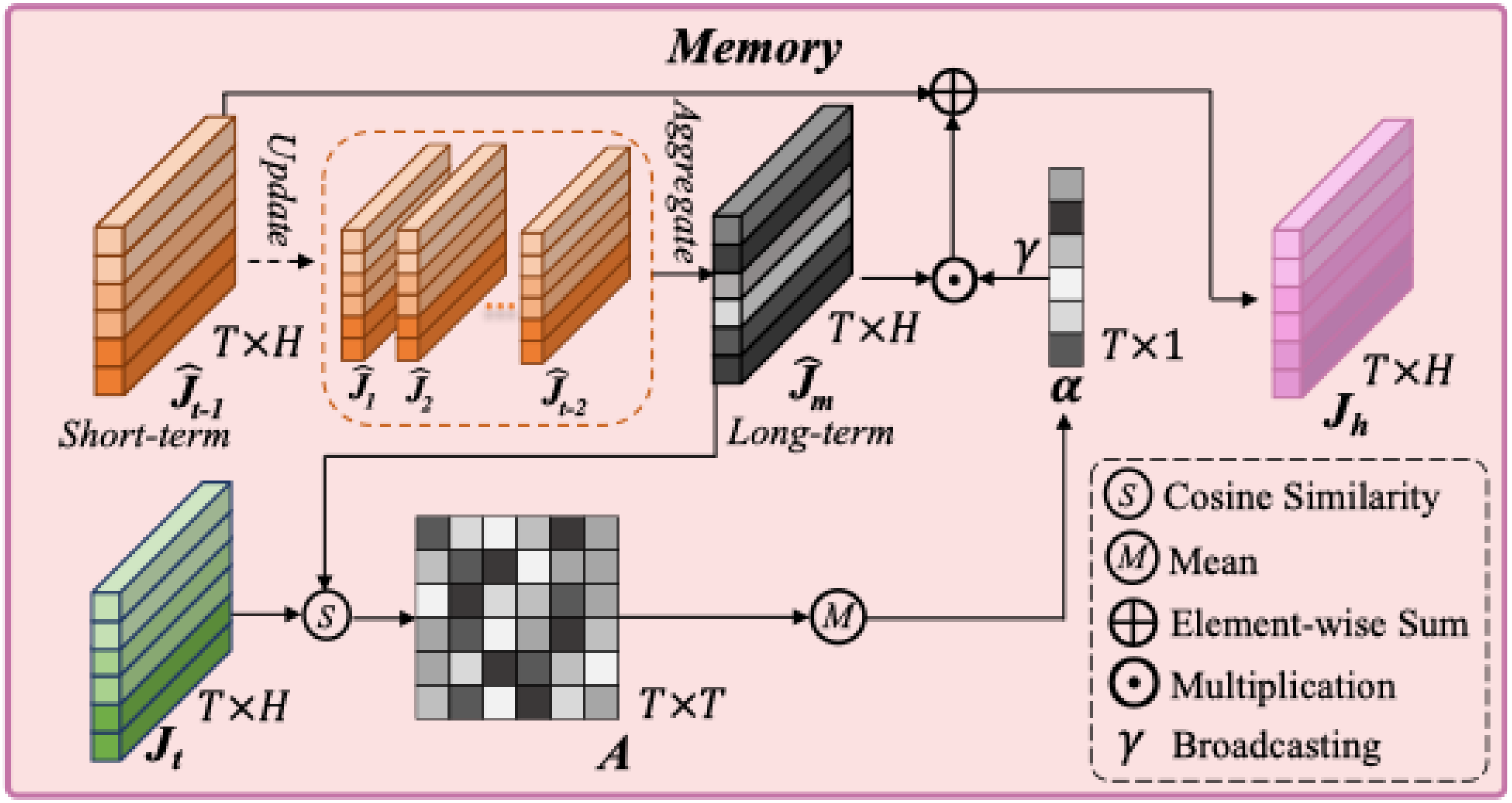
GroundFlow



$$\begin{cases} \hat{J}_t = J_t + \text{softmax}\left(\frac{J_t J_h^T}{\sqrt{H}}\right) J_h, & t \in [2, n] \\ \hat{J}_t = J_t, & t = 1 \end{cases},$$

Inject critical history information J_h into current step embedding J_t

GroundFlow



Sum-up history information

$$\hat{J}_m = \sum_{t=1}^{t-2} \hat{J}_t,$$

Calculate importance based on current need

$$\mathbf{A}_{i,j} = \circledS\left(\hat{J}_m, J_t\right),$$

Average

$$\alpha = \frac{1}{T}\sum_{j=1}^{T} \mathbf{A}_{i,j},$$

Short-term + weighted long-term

$$J_h = \hat{J}_{t-1} + \hat{J}_m \odot \gamma(\alpha),$$

Experiments

Model Type	Method	ScanNet		3RScan		MultiScan		ARKitScenes		HM3D		Overall	
		s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc	s-acc	t-acc
LLM-based	GPT4 + PointNet++ (Zero-shot) [55]	42.6	10.9	25.5	2.4	27.0	0.0	27.6	6.0	20.8	7.7	27.3	7.6
	LEO (3DLLM) [25]	61.2	25.7	55.8	16.0	52.7	7.6	69.6	41.5	61.5	35.7	62.8	34.1
Dual-stream	3D-VisTA [59]	60.1	24.7	52.7	13.5	47.6	7.0	68.4	37.8	57.5	30.6	60.3	28.8
	3D-VisTA+ **GroundFlow**	63.0	26.6	56.8	21.7	57.1	14.0	71.9	46.0	62.3	36.9	64.1	35.1
	MiKASA [6]	57.8	19.4	53.0	10.9	48.7	2.3	67.1	35.7	57.3	30.1	60.8	31.9
	MiKASA + **GroundFlow**	62.7	**28.9**	58.9	17.4	54.0	11.6	70.2	42.9	61.8	36.2	63.5	34.2
Query-based	PQ3D [60]	53.7	17.9	50.2	9.9	43.5	4.7	64.9	32.0	56.9	30.6	57.3	25.9
	PQ3D + **GroundFlow**	62.0	28.2	**60.1**	21.0	51.3	7.0	**73.0**	**48.1**	**63.6**	**38.0**	**64.8**	**36.1**
	Vil3DRel [9]	59.3	19.9	55.9	15.2	50.9	4.7	69.3	38.6	58.7	31.0	61.1	28.6
	Vil3DRel + **GroundFlow**	**63.1**	27.8	58.8	**22.5**	**57.6**	**20.9**	72.4	45.1	62.3	36.6	64.4	35.2

- Significant performance improvements (1.2 ~ 2.1 times in t-acc) can be observed when 3DVG baselines integrate with GroundFlow.
- Achieve SOTA performance, out-perform fine-tuned 3DLLM LEO.

Inference Speed

Models	#params	Speed	s-acc	t-acc
LEO	6.9B	11.3ms	62.8	34.1
3D-VisTA	101.1M	5.2ms	60.3	28.8
3D-VisTA+ GroundFlow	**123.1M**	**5.6ms**	**64.1**	**35.1**
PQ3D	167.4M	6.8ms	57.3	25.9
PQ3D+ GroundFlow	**189.4M**	**6.9ms**	**64.8**	**36.1**

Table 1

Dataset	Scan	3R	Multi	ARK	HM	Overall
Avg. Scene Size	30.7	31.5	40.8	12.1	31.0	25.1
Avg. Speed (ms)	5.71	5.78	6.66	5.20	6.13	5.63

Table 2

Ablation Study

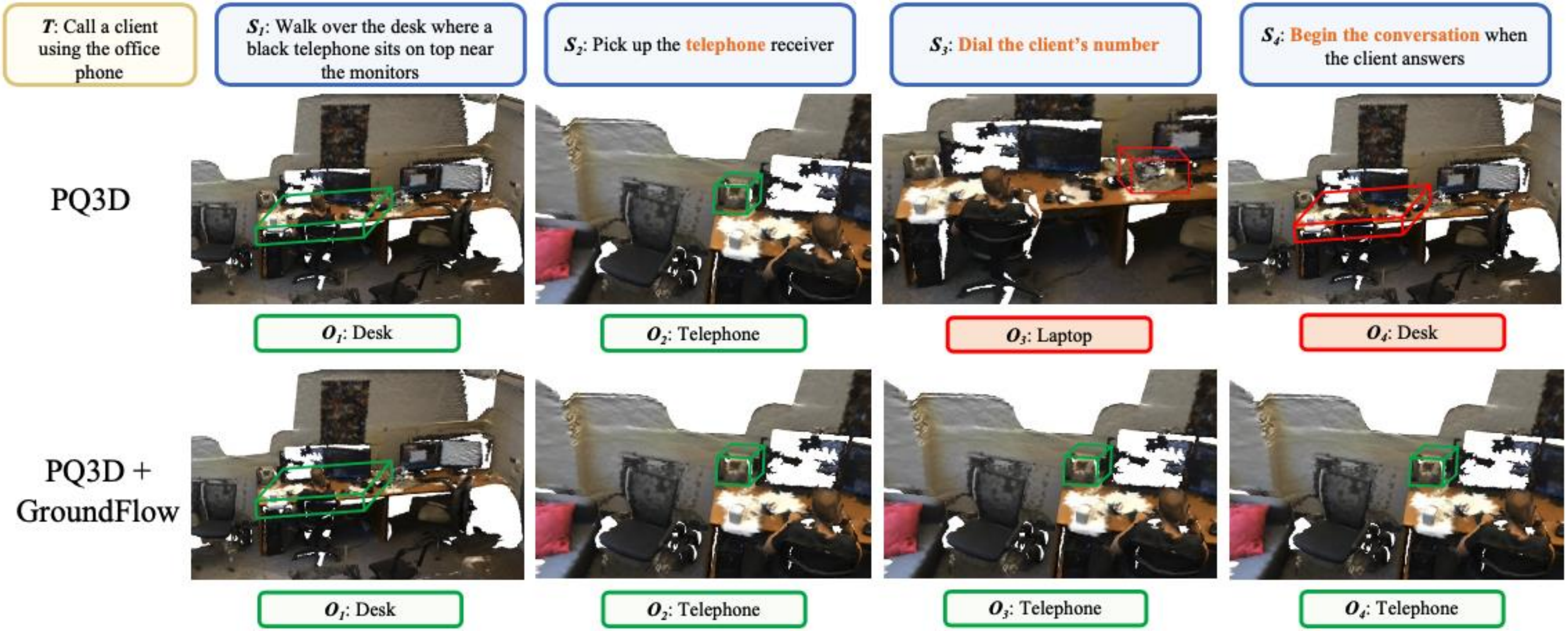
Models	Temporal Fusion Methods	s-acc	t-acc	Δs-acc	Δt-acc
3D-VisTA	LSTM	61.4	29.5	+1.1	+0.7
	GRU	62.0	28.8	+1.7	+0.0
	Transformer	62.9	33.5	+2.6	+4.7
	GroundFlow	**64.1**	**35.1**	**+3.8**	**+6.3**
PQ3D	LSTM	63.1	30.8	+5.8	+4.9
	GRU	63.8	30.7	+6.5	+4.8
	Transformer	63.4	33.6	+6.1	+7.7
	GroundFlow	**64.8**	**36.1**	**+7.5**	**+10.2**

Table 3

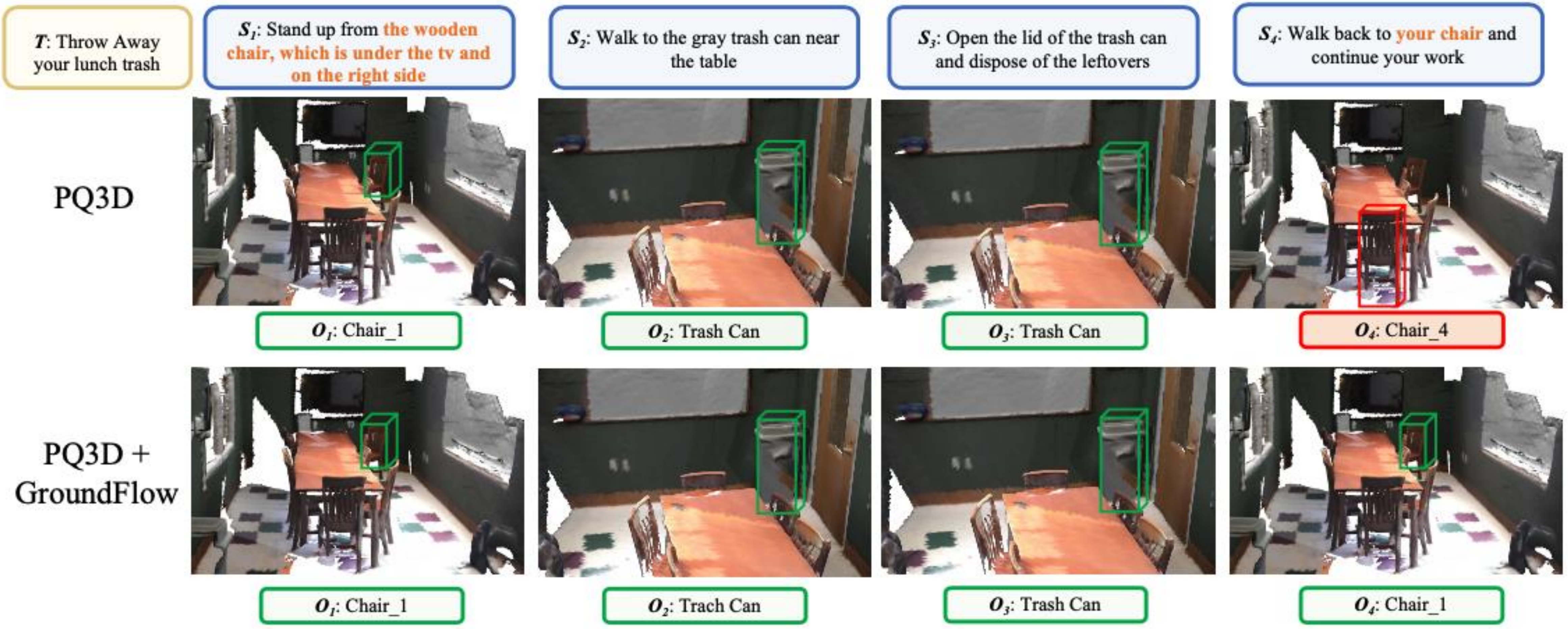
Step/Dataset	Scan	3R	Multi	ARK	HM	Overall
2	+26.3	/	/	/	+4.8	+7.3
3	+18.5	+6.7	/	+12.1	+7.7	+10.7
4	+8.2	+10.3	/	+10.7	+1.9	+6.0
5	+4.4	+10.0	/	+11.3	+6.4	+7.6
6	+9.7	+7.8	/	+14.5	+9.0	+10.4
7	+3.3	+12.5	/	+21.9	+8.7	+11.7
≥ 8	+14.3	/	/	+18.4	+8.4	+10.6

Table 4

Qualitative Visualization



Qualitative Visualization



Conclusions

- Lightweight (22M parameters) plug-in designed for 3DVG baselines, capable of effectively extracting relevant short-term and long-term instructions, enabling a smooth transition to challenging SG3D tasks.

- Evaluated in SG3D benchmark, GroundFlow results in significant improvements for t-acc and s-acc without affecting inference speed.

- 3DVG + GroundFlow achieves SOTA performance in SG3D benchmark, surpassing fine-tuned 3DLLM.

Thank you!