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

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

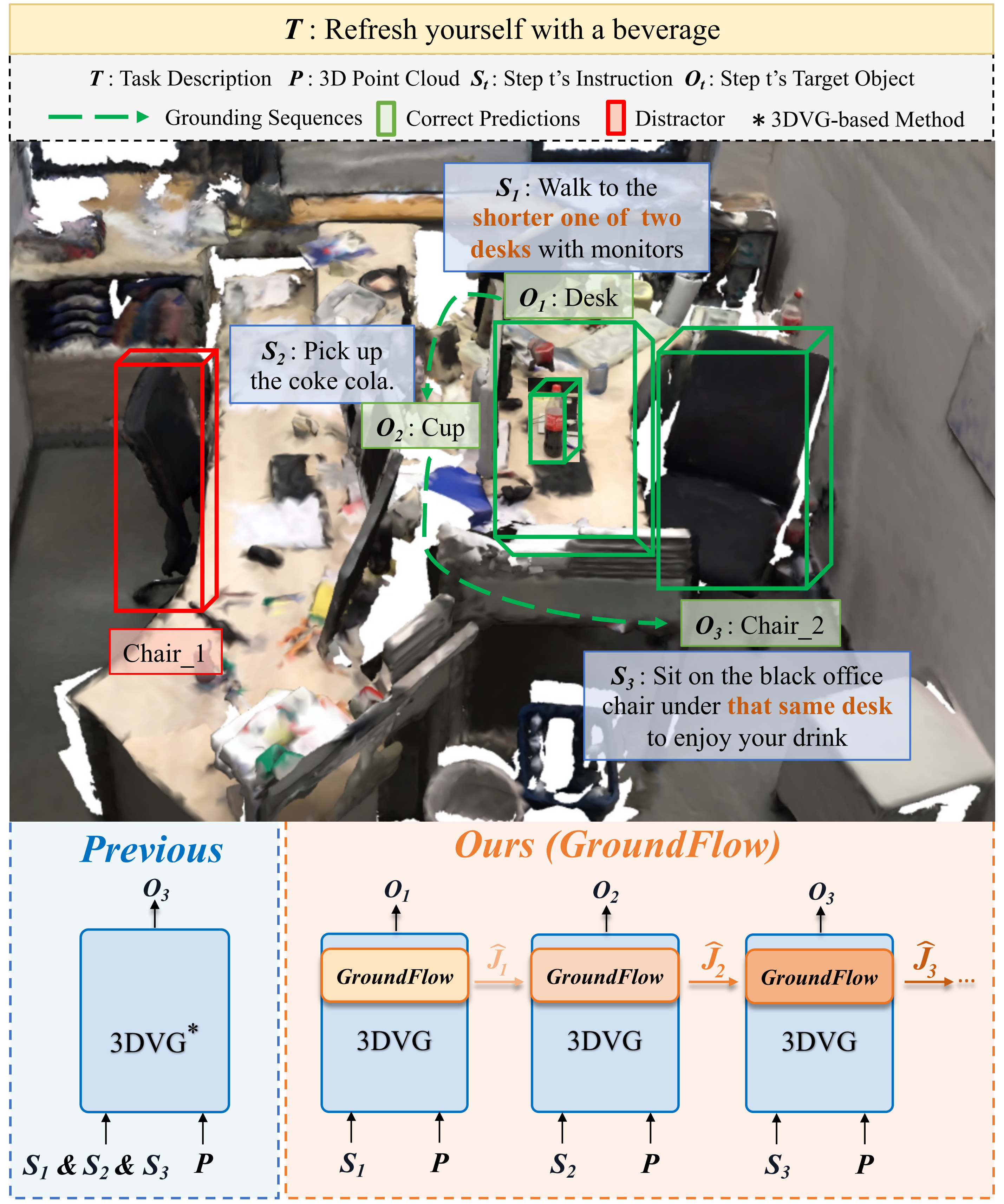
Nanyang Technological University[1], Centre for Frontier AI Research, A*STAR[2], Shanghai University of Finance and Economics[3]

Paper

Code

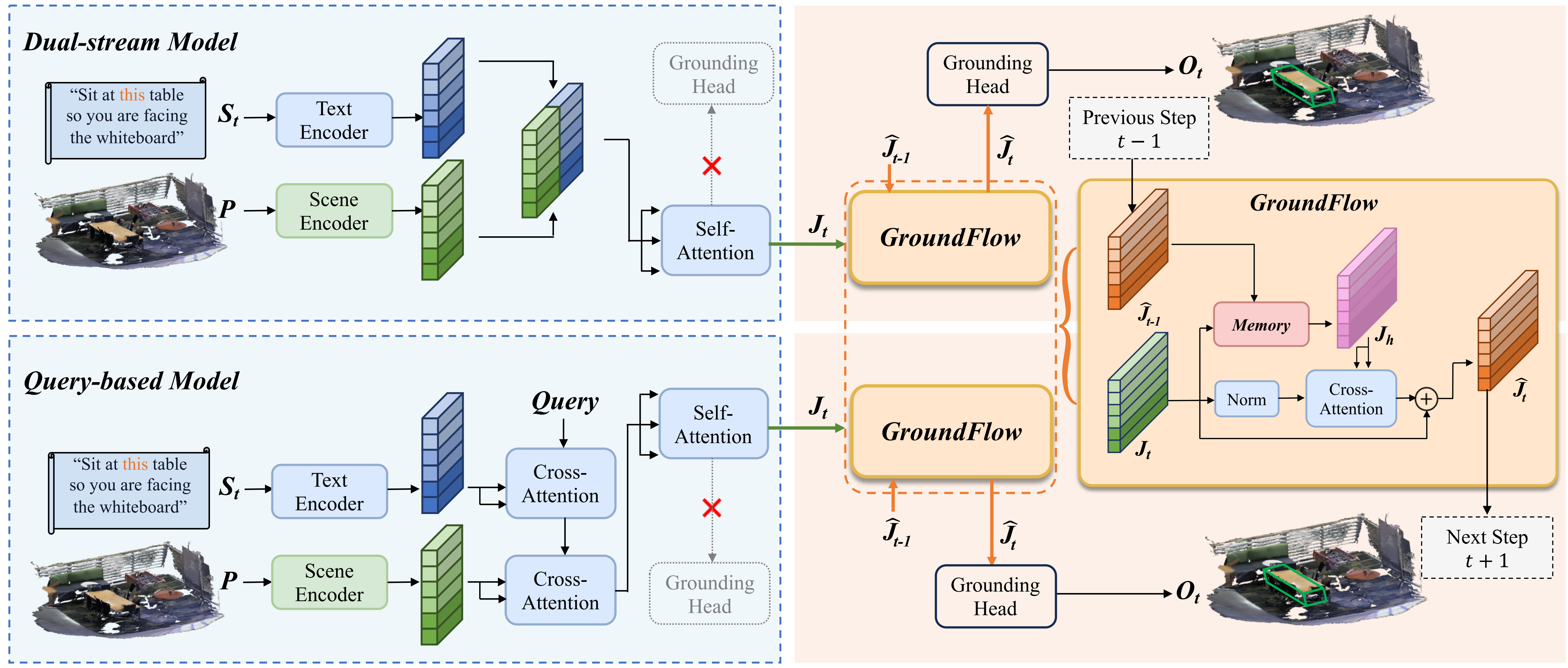
ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation


T : Refresh yourself with a beverage
T : Task Description P : 3D Point Cloud S_t : Step t's Instruction O_t : Step t's Target Object
Grounding Sequences Correct Predictions Distractor * 3DVG-based Method
S_1 : Walk to the shorter one of two desks with monitors
O_1 : Desk
S_2 : Pick up the coke cola.
O_2 : Cup
O_3 : Chair_2
Chair_1
S_3 : Sit on the black office chair under that same desk to enjoy your drink
Previous
O_3
3DVG*
S_1 & S_2 & S_3 P
Ours (GroundFlow)
O_1 O_2 O_3
GroundFlow GroundFlow GroundFlow
3DVG 3DVG 3DVG
$\hat{J}_1$ $\hat{J}_2$ $\hat{J}_3$ ...
S_1 P S_2 P S_3 P

- SG3D task instruction often contains **pronouns** such as "it", "the other", "the same".
- It requires grounding method to understand the **context** and retrieve **relevant history information.**
- Previous 3DVG method: Simply **concatenate** multiple step instructions **without extracting temporal information.**
- We propose GroundFlow – a **plug-in** module for **temporal reasoning** on SG3D.

GroundFlow



☑ GroundFlow could be built on any 3DVG models as a plug-in in a recurrent framework.

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) [52]	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) [24]	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 [56]	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 [5]	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 [57]	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 [8]	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

☑ GroundFlow improves the task accuracy of baseline methods by large margin (**+7.5%** in dual-stream and **+10.2%** in query-based).

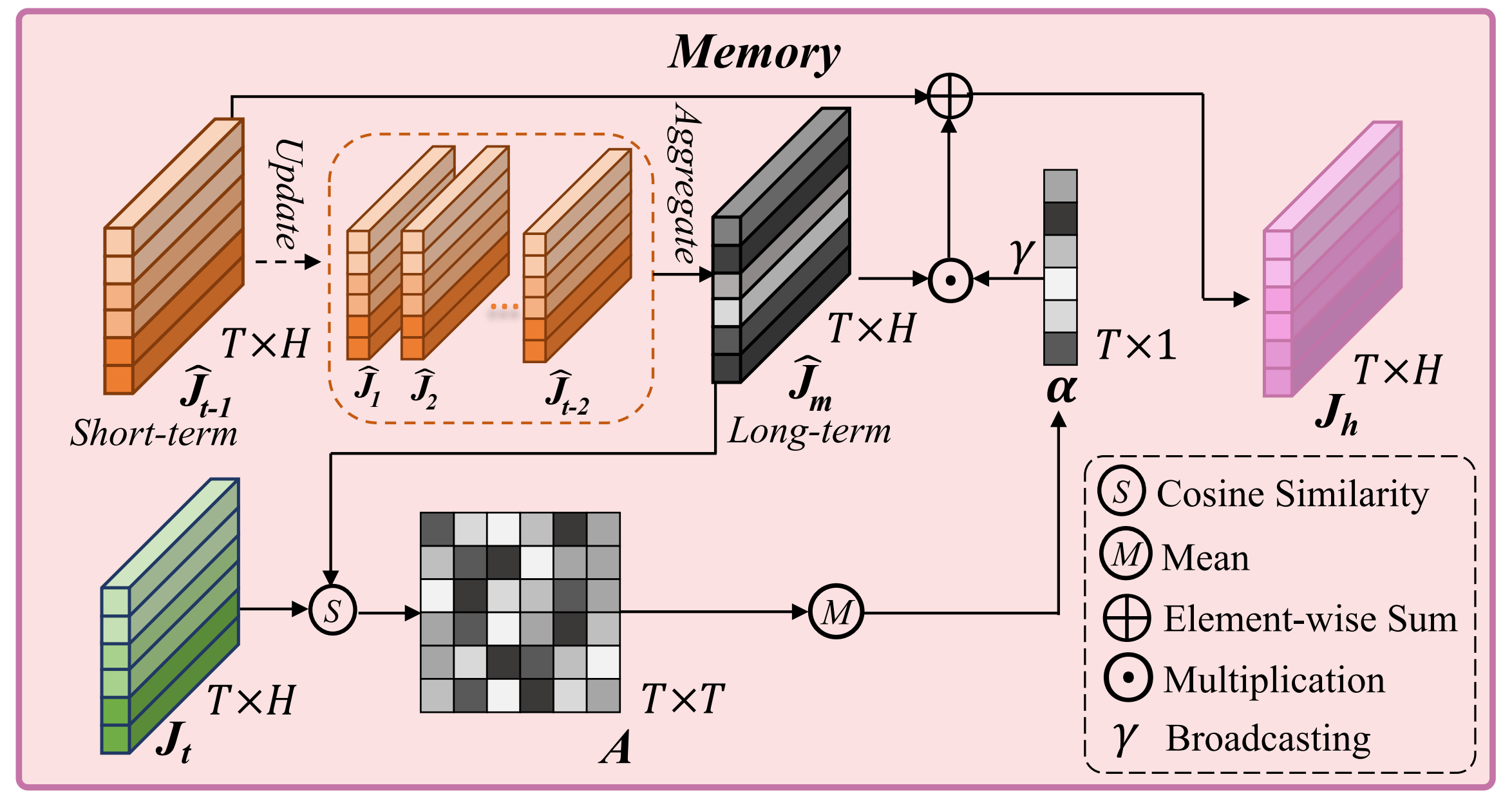
☑ 3DVG + GroundFlow outperforms 3DLLM LEO, achieving **SOTA** results.

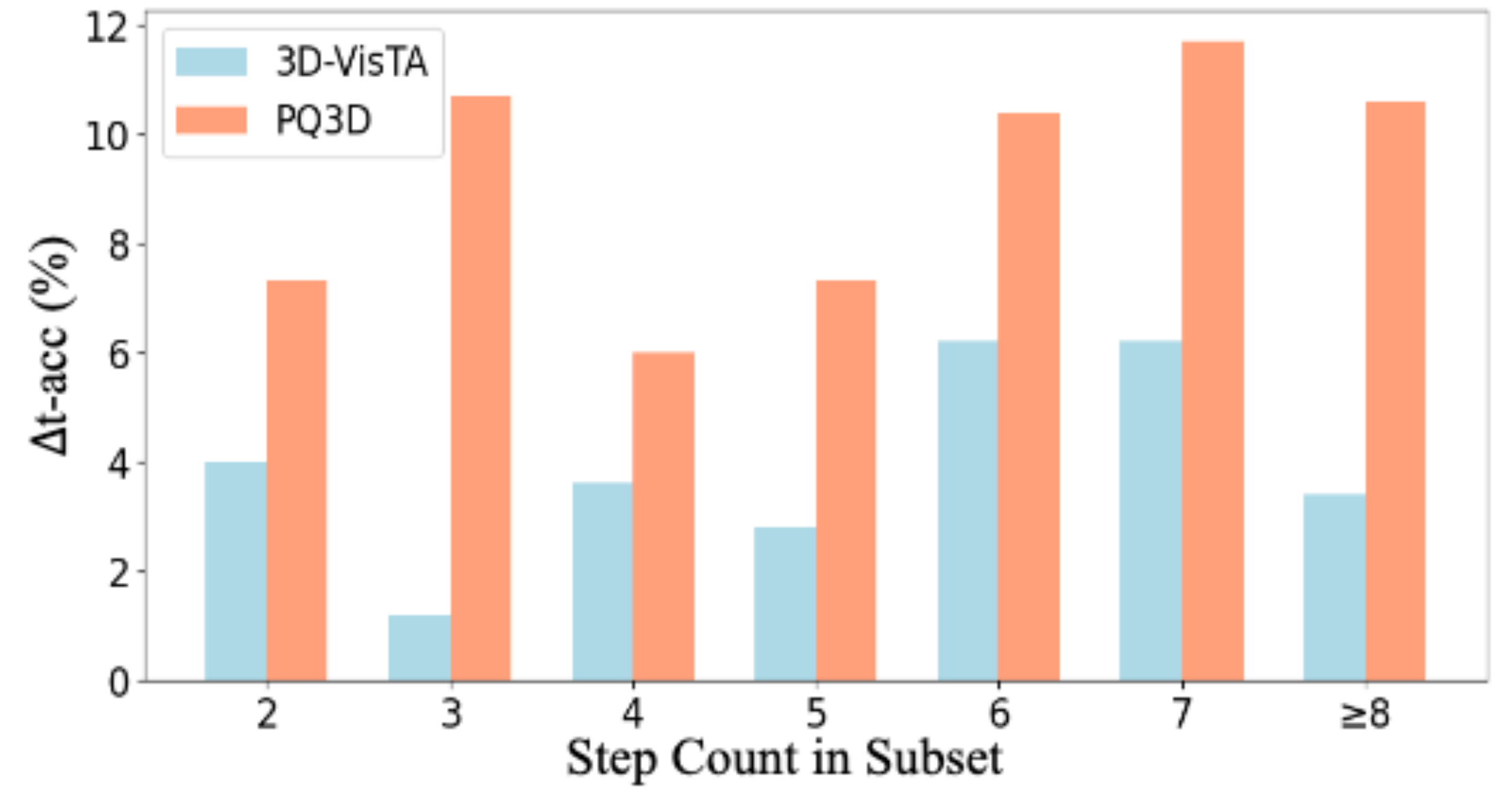
Models	Temporal Fusion Methods	s-acc	t-acc	Δs-acc	Δt-acc
3D-VisTA	LSTM	61.4	29.5	+1.1	+0.7
	GRU	62.0	28.8	+1.7	+0.0
	Transformer	62.9	33.5	+2.6	+4.7
	GroundFlow	**64.1**	**35.1**	**+3.8**	**+6.3**
PQ3D	LSTM	63.1	30.8	+5.8	+4.9
	GRU	63.8	30.7	+6.5	+4.8
	Transformer	63.4	33.6	+6.1	+7.7
	GroundFlow	**64.8**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

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**

☑ GroundFlow performs temporal fusion more effectively than traditional methods (LSTM, GRU or Transformer) with only **22M** parameters and a **marginal increase in inference time.**

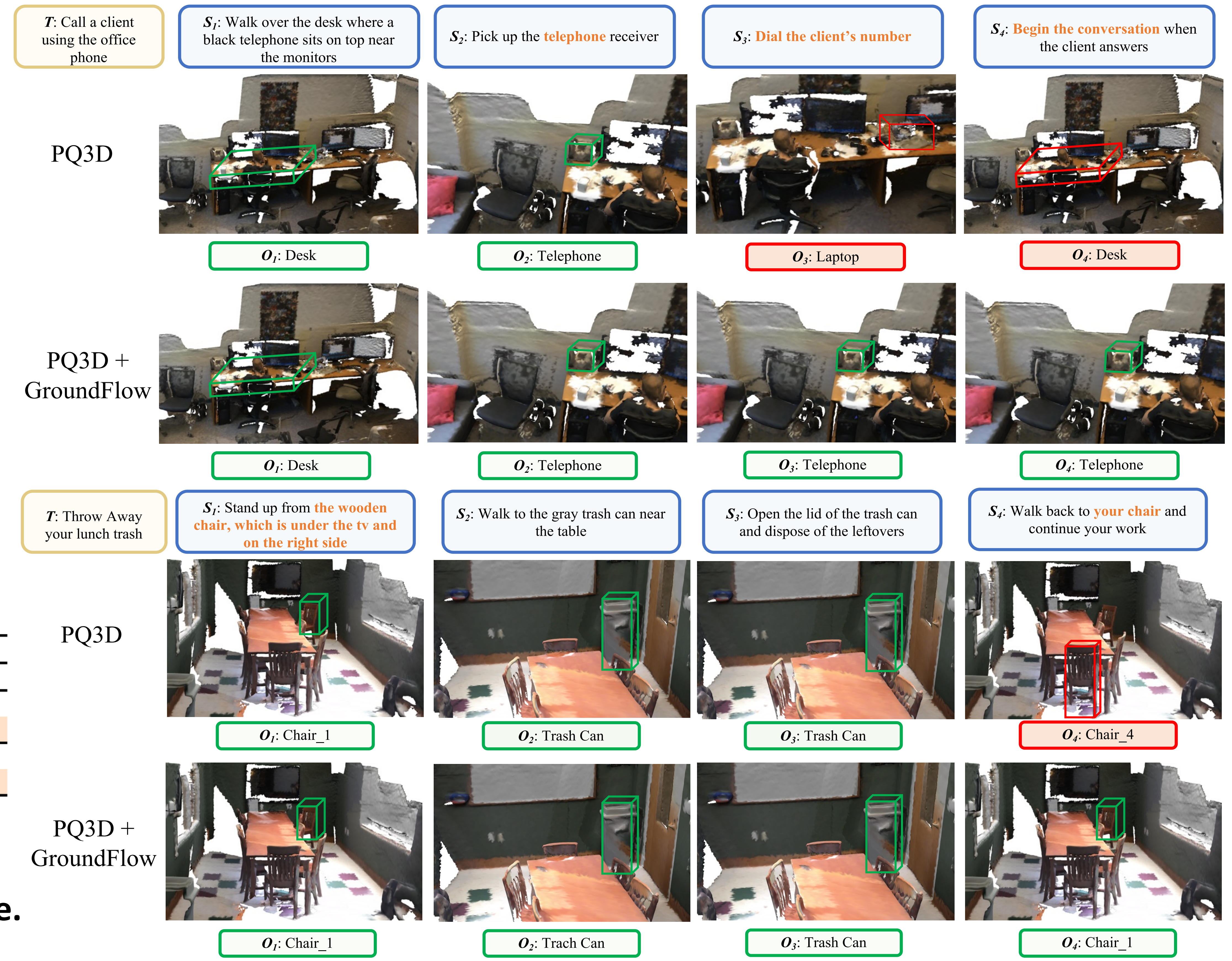
Memory Structure





☑ GroundFlow **selectively** extract short-term and long-term information **based on its relevance to the current instruction**, maintaining its temporal understanding advantage as **step counts increase**.

Qualitative Visualization


T: Call a client using the office phone
S_1: Walk over the desk where a black telephone sits on top near the monitors
S_2: Pick up the telephone receiver
S_3: Dial the client's number
S_4: Begin the conversation when the client answers
PQ3D
O_1: Desk
O_2: Telephone
O_3: Laptop
O_4: Desk
PQ3D + GroundFlow
O_1: Desk
O_2: Telephone
O_3: Telephone
O_4: Telephone
T: Throw Away your lunch trash
S_1: Stand up from the wooden chair, which is under the tv and on the right side
S_2: Walk to the gray trash can near the table
S_3: Open the lid of the trash can and dispose of the leftovers
S_4: Walk back to your chair and continue your work
PQ3D
O_1: Chair_1
O_2: Trash Can
O_3: Trash Can
O_4: Chair_4
PQ3D + GroundFlow
O_1: Chair_1
O_2: Trach Can
O_3: Trash Can
O_4: Chair_1