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

Supplementary Material

A. SG3D Benchmark Statistics

In Table 4, we show the detailed dataset statistics in the SG3D benchmark. There are 22,346 tasks and 112,336 steps in total, with an average of 5.03 steps per task, an average length of 12.7 words per step instruction, and 70.5 words per task. All five datasets in SG3D were split into training and evaluation sets.

Dataset	$\#P$	$\#O/P$	$\#T$	$\#S$
ScanNet	693	30.7	3,174	15,742
3RScan	472	31.5	2,194	11,318
MultiScan	117	40.8	547	2,683
ARKitScenes	1,575	12.1	7,395	39,887
HM3D	2,038	31.0	9,036	42,706
Total	4,895	25.1	22,346	112,336

Table 4. Dataset statistics of SG3D benchmark. $\#P$, $\#O/P$, $\#T$ and $\#S$ represent the number of 3D point cloud scenes, the average number of candidate objects per scene, the number of tasks and the number of steps, respectively.

B. Model Computational Complexity

In Table 5, 3D-VisTA and PQ3D are selected to integrate with the GroundFlow module as two examples to demonstrate a better trade-off between accuracy and speed. With only a marginal increase in inference time (approximately 0.4ms for 3D-VisTA and 0.1ms for PQ3D), both step accuracy and task accuracy improve significantly. Additionally, since GroundFlow has only 22M parameters, all 3DVG + GroundFlow experiments can be efficiently deployed on a single NVIDIA 24GB A5000 GPU.

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 5. Comparisons of size, speed and performance of different models. #params indicates the number of parameters each model has and speed shows the inference time per step.

C. Data efficiency

Figure 6 shows that increasing the amount of data improves the performance of all methods. Notably, 3DVG methods integrated with GroundFlow demonstrate superior data efficiency. They achieve performance comparable to 3DVG baselines using only 50% of the data and surpass the 3DLLM LEO model with less than 75% of the data. This advantage likely stems from GroundFlow's specialized design for sequential grounding task, which enables the model to efficiently learn from historical information in context and generalize to unseen step instructions.

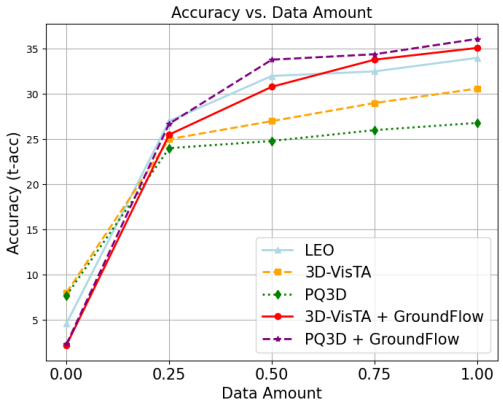

Figure 6. Comparisons of data amount versus task accuracy (t-acc).

D. Failure Cases

We found that most failures of GroundFlow occur when an incorrect prediction is made in previous steps. As shown in Figure 7, since GroundFlow is built on 3DVG methods, any incorrect information generated by 3DVG methods gets propagated to future steps, leading to wrong predictions.

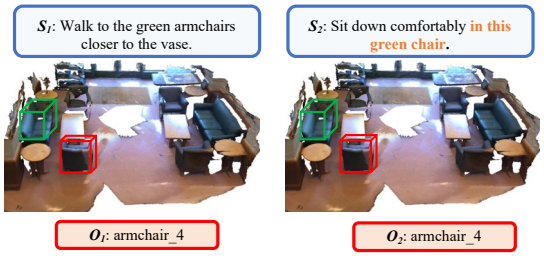

Figure 7. Example of the failure case.

E. More Qualitative Comparisons

We include more qualitative comparisons in Figure 8. It is shown that GroundFlow enables the 3DVG baseline models to effectively capture contextual past information to make current step predictions.

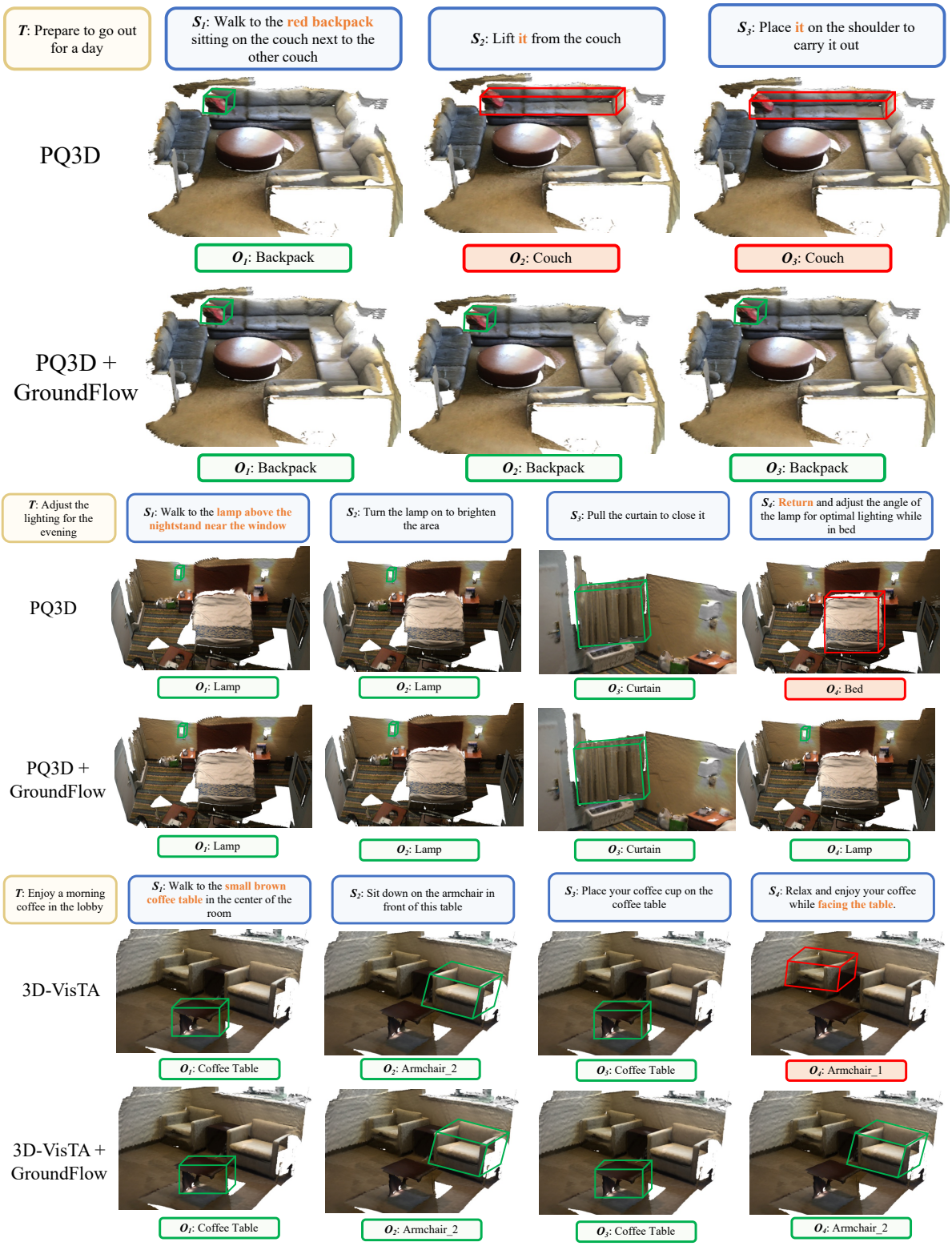

Figure 8. Visualization results from PQ3D and 3D-VisTA without and with GroundFlow. T represents the task description, S_t and O_t denote the step instruction and corresponding referred target object in step t. Red are wrong predictions and green are correct predictions.