

Advancing 3D Object Grounding Beyond a Single 3D Scene – Supplementary Material –

Section A of the supplementary material provides more detailed statistics of the proposed G-Sr3D-ST/MT datasets. In Section B, we study the effectiveness of multi-level grounding and discuss the computational efficiency of our model. In Section C, we provide additional visualization results and qualitative analysis.

A DETAILED STATISTICS OF G-SR3D-ST/MT

Table 1: G-Sr3D-ST/MT statistics on train/val/test splits.

Dataset	Split	1	2	3	4	5	6	7	8
G-Sr3D-ST	Train	15467	7455	6015	4987	3791	3202	2857	2518
	Val	3921	1890	1525	1264	962	812	724	638
	Test	2396	1155	932	773	587	496	443	390
G-Sr3D-MT	Train	11731	7421	5042	2972	2334	2104	1840	1751
	Val	3519	2226	1513	892	700	631	552	525
	Test	1508	954	648	382	300	270	236	225

In Tab. 1, we show the number of sentence-scene group pairs that contain different numbers (1 to 8) of positive scenes across the train/val/test split in the G-Sr3D-ST/MT datasets, where the total numbers are 46292, 11736, 7172, and 35195, 10558, 4523, respectively. The groups containing 2 to 4 positive scenes account for more than half of all multi-scene groups, and the groups containing 5 to 8 positive scenes are roughly evenly distributed. Cases of different positive scene numbers all have sufficient samples for training, validation, and testing. We visualize the distribution of the number of sentence-scene group pairs and the average number of target objects per sentence broken down by positive scene number and object type in Fig. 1. We show these statistics for the 15 most frequently referred object types in the G-Sr3D-ST/MT datasets. We see that the Single Target sentence-scene group pairs reflect the

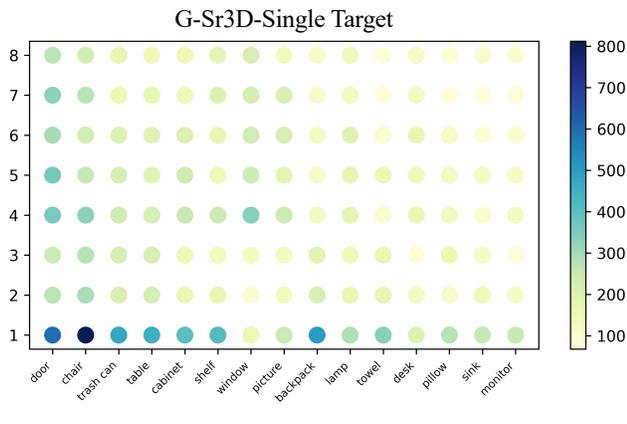


Table 2: Ablation study of Multi-level Grounding in GNL3D.

Method	ST w/ Negatives			MT w/ Negatives		
	Single	Multiple	All	Single	Multiple	All
w/o Multi-level Grounding	72.0	78.2	75.5	61.9	62.4	62.2
w/ Multi-level Grounding	72.8	79.1	76.4	63.1	63.6	63.4

distributions of objects in the real world. For the Multiple Targets data pairs, there are more groups containing 1 to 3 positive scenes and there are 3.26 targets per sentence on average. Chairs, pillows, and books are the 3 most common object types with the number of targets ranging from 2 to 7.

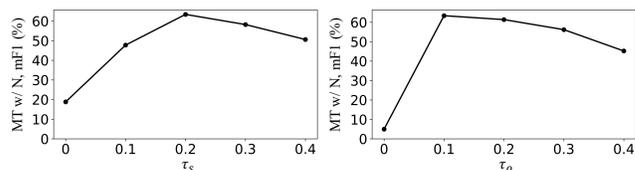


Figure 2: Effect of different values of τ_s and τ_o .

B ADDITIONAL EXPERIMENTS

Ablation study of multi-level grounding. In our GNL3D, we design two types of grounding heads to perform 3D grounding at both scene-level and object-level. We study the effect of this multi-level grounding strategy on the G-Sr3D-ST/MT with Negatives datasets in Tab. 2. Note that “w/o Multi-level Grounding” denotes the model without the scene-level grounding head, which makes the scene- and object-level predictions based solely on the object-level

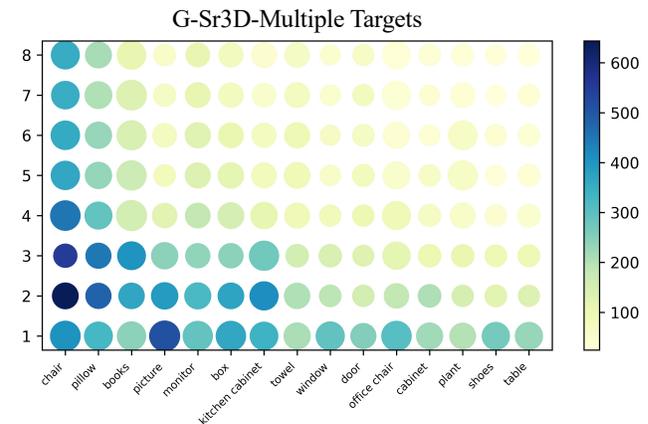


Figure 1: Distribution of the number of sentence-scene group pairs (color) and the average number of target objects per sentence (circle size) by positive scene number and object type for the 15 most frequently referred object types on G-Sr3D-ST/MT.

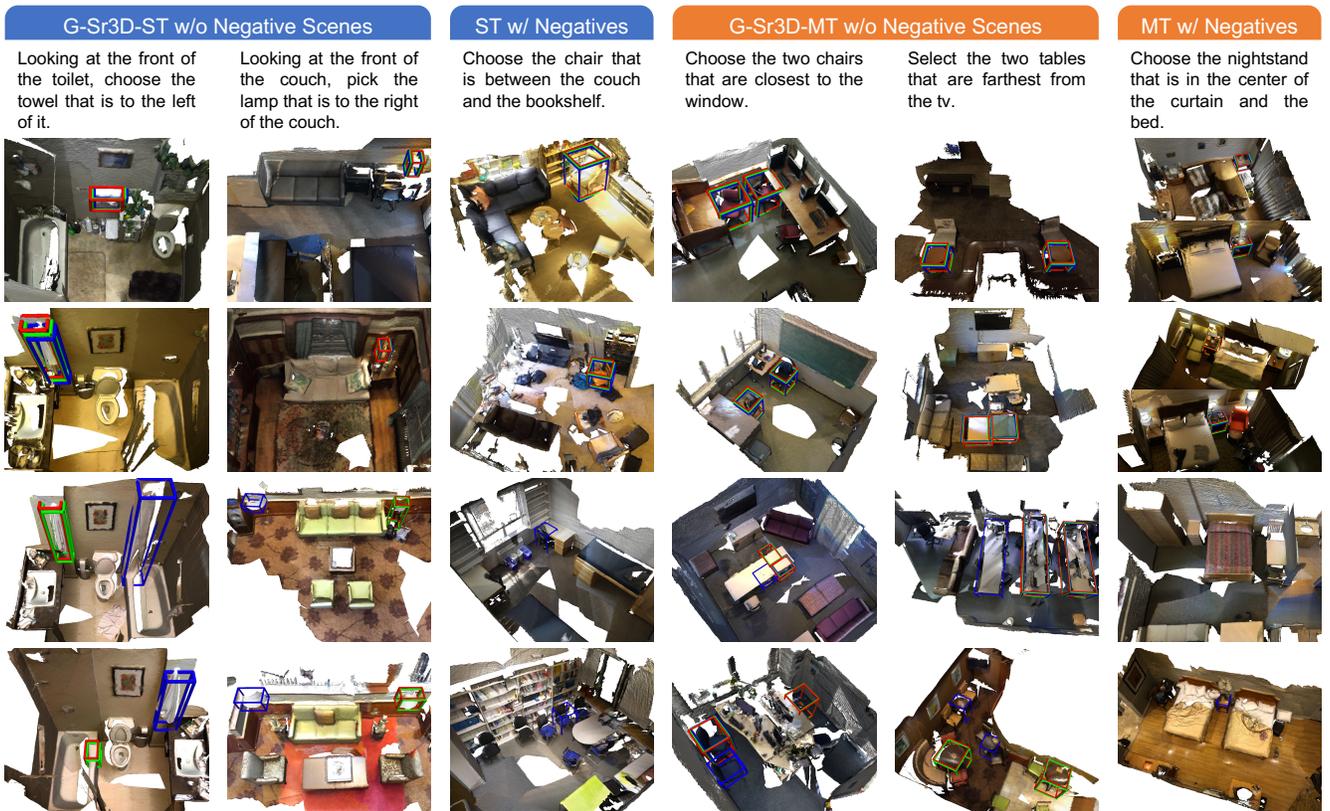


Figure 3: Qualitative comparison results on the Group-wise 3D Object Grounding task. Green boxes indicate Ground Truth, blue boxes are 3D-VisTA’s results, and red boxes are our GNL3D’s results.

grounding head. In the single-target setting, the object proposal with maximum probability is taken as the target object only if its probability is above a threshold τ_1 . In the multi-target setting, all object proposals with predicted scores above a threshold τ_2 are predicted as target objects. Scenes without predicted target objects are taken as negative scenes. We see that our GNL3D with multi-level grounding outperforms the model with a single object-level grounding head, validating the effectiveness of our design. We further study different τ_s and τ_o used in the scene- and object-level grounding heads to filter our model outputs on G-Sr3D-MT with Negatives. Fig. 2 shows that when $\tau_s = 0.2$ and $\tau_o = 0.1$, our GNL3D achieves the best performance.

Computational efficiency. We compare the floating-point operations per second (FLOPs), model parameters, and the overall mean accuracy (mAcc) with the base model 3D-VisTA. For the FLOPs and model parameters, we re-implement 3D-VisTA with 6 scene encoding layers for a fair comparison and use the same input with batch size 1 and group size 8. As for the mAcc, we report the overall results on G-Sr3D-ST without Negatives. As shown in Tab. 3, our GNL3D significantly outperforms the base model while introducing only 16% additional parameters and less than 1% FLOPs overhead. This is because our introduced LCAM is not the computational bottleneck since its computational complexity is $O(KM \cdot T)$, which

Table 3: Comparison of flops, model parameters, and overall mean accuracy between 3D-VisTA and GNL3D on ST w/o N.

Method	FLOPs (G)	Params (M)	mAcc (%)
3D-VisTA	20.59	78.41	69.9
GNL3D (Ours)	20.72	91.03	74.6

is linearly related to the group size K , the average number of objects per 3D scene M , and the number of tokens per sentence T .

C QUALITATIVE ANALYSIS

We provide additional qualitative results in this section. We qualitatively compare GNL3D to the baseline model 3D-VisTA on G-Sr3D-ST/MT in different group-wise scenarios and display some typical examples in Figure 3. By intuitive comparison, our GNL3D gives more precise group-wise grounding results than 3D-VisTA. Moreover, our method can better understand the described object’s category and spatial relations to anchor objects. These examples demonstrate that our proposed LCAM mechanism can explicitly exploit the intra-group vision-vision connections to build a more accurate target concept for GNL3D.