Multi3DRefer: Grounding Text Description to Multiple 3D Objects Supplemental Materials

Yiming Zhang[1] ZeMing Gong[1] Angel X. Chang[1,2]
Simon Fraser University[1] Alberta Machine Intelligence Institute (Amii)[2]
{yza440, zmgong, angelx}@sfu.ca
https://3dlg-hcvc.github.io/multi3drefer/

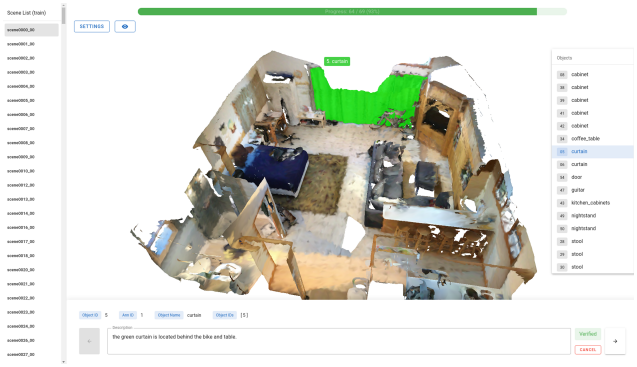

Figure 1: A screenshot of the Multi3DRefer verification web interface. Verifiers use the interative 3D interface to check that the description matches the selected objects, and modify the list of selected objects or description if needed.

(a) The target stool is to the right of the bed.

(b) The target chair is to the right of the table.

(c) The chair is located to the right of you, as you stand on the black rug and face the coffee table.

Figure 2: Examples of descriptions with spatial relation show to annotators with target objects to illustrate when (a) a single object matches, (b) multiple objects can match depending where the viewer would be, and (c) a case where the viewpoint is specified in the description.

In this supplement, we provide more details about the web interface for verification (Appendix A), and statistics (Appendix B) and qualitative examples from our dataset (Fig. 3). We also present a discussion of the computational efficiency of our online renderer (Appendix C), analysis of matching strategies and thresholds on the Multi3DRefer task (Appendix D), and provide additional qualitative examples of our M3DRef-CLIP (Appendix E).

A. Web interface for verification

We implement a web-based data verification application using Three.js[1], Vue.js[2] and FastAPI[3], to allow human verifiers to verify and correct the generated data. See Fig. 1 for a screenshot of our web interface. Verifiers are shown a generated description together with an interactive 3D mesh of a scene, where the selected objects are highlighted in green. Verifiers are asked to check whether the description matches the identified target objects (in green). If the description does not match, verifiers are asked to either: 1) change the target object list (by clicking on objects in the scene to toggle selection); or 2) modify the description if necessary. Once the description clearly matches the selected objects, the 'Verify' button is clicked to indicate that the pair has been manually verified. Verifiers are instructed to consider whether a viewpoint is specified in the description or not. If a specific viewpoint is specified, then the viewpoint should be used to identify the specific objects being described. If no viewpoint is specified, then annotators are instructed to imagine different potential viewpoints from which they can stand and all objects that can match the given description. See Fig. 2 for examples shown to annotators.

In total, verifiers checked 64513 description-scene pairs. Of these, we discard 2587 samples (542 from Zero Target and 2045 from Multiple Targets) to limit the number of zero-target descriptions per scene to 21 and the number of overly similar descriptions for complex scenes. During verification, 11804 descriptions were modified by verifiers. Most of the modifications were minor changes such as changing 'left' to 'right', or adding more constraints such as

[1]https://threejs.org/
[2]https://vuejs.org/
[3]https://fastapi.tiangolo.com/

	Train	Val	Test	**Total**
#descriptions	43,838	11,120	6,968	61,926
#scenes	562	141	97	800
#objects	8,346	2,161	1,102	11,609
avg. #objects / scene	14.9	15.3	11.4	14.5
avg. #descriptions / scene	78.0	78.9	71.8	77.4
avg. #descriptions / object	5.3	5.1	6.3	5.3

Table 1: Multi3DRefer dataset statistics. on Train, Val and Test sets.

	Training		Inference	
	Mem	Time	Mem	Time
D3Net (Grounding)	14.7G	41.1m	15.2G	10.1m
M3DRef-CLIP	15.2G	55.5m	11.3G	12.5m

Table 2: Comparison of GPU memory usage and running time between D3Net [1] and M3DRef-CLIP.

enhancing 'this is a chair' to 'this is a chair facing the wall'. The verification check took about 9 seconds per zero target description and 16 seconds per Single Target / Multiple Targets description.

B. Statistics and examples of Multi3DRefer

Tab. 1 show detailed statistics and Fig. 3 show examples of the Multi3DRefer dataset with descriptions matching zero, single, or multiple targets.

C. Computational efficiency

We compare training and inference time and GPU memory usage with the D3Net [1] grounding module (which also uses PointGroup [2] as the detector), using *torch.cuda.max_memory_reserved* (Tab. 2). We use the same input and batch size 4 for 60 epochs until convergence and report GPU memory and time per epoch for the same machine with an NVIDIA RTX A5000 GPU. The memory and computation overhead is only 10-20%, including all rendering.

D. Analysis of matching strategies

We study the effect of different matching strategies (*Hungarian* vs *All*) together with matching IOU thresholds τ_{train} for training, and different prediction confidences on the performance of the D3Net and M3DRef-CLIP on the Multi3DRefer task. We plot the F1 at IOU of 0.5 for these different variants using the 5-scenarios breakdown we established.

5-scenario breakdown. We identify our 5 scenarios (ZT w/ D, ZT w/o D, ST w/ D, ST w/o D and MT) according to the nyu40 semantic label set. Due to our metrics setting, ZT metrics are special cases which report $F1 = 1$ if the model predicts nothing and $F1 = 0$ if the model predicts too many. (Fig. 4)

Prediction threshold. We further study different τ_{pred} used to filter out model outputs. Fig. 4 shows that all models achieve the best performance at $\tau_{\text{pred}} = 0.1$.

Matching strategies. We compare the two matching strategies (*Hungarian* vs *All*) that we used to set up positive and negative instances between the GT bounding boxes and proposed bounding boxes for calculating the reference loss L_{ref}. We compare results between M3DRef-CLIP and D3Net [1]. Fig. 4 shows that *Hungarian* (darker lines) outperforms *All* (lighter lines) on both methods, especially when τ_{train} is small (e.g. 0.25), since *Hungarian* guarantees an optimal one-to-one matching. When τ_{train} is larger (e.g. 0.5), the gap caused by these two strategies gradually narrows. For D3Net [1], the two matching strategies does not exhibit a noticeable differences. We suspect this is due to a less noisy detector and that *Hungarian* matching is effective when proposals are noisy.

E. Qualitative results on Multi3DRefer

In Fig. 5, we show qualitative examples of outputs from D3Net [1] and M3DRef-CLIP for zero, single, and multiple targets. In the Zero Target case (column 1), M3DRef-CLIP tends to predict false positives. In the Single Target case (column 2), M3DRef-CLIP has more accurate bounding boxes. For Multiple Targets case (column 3), M3DRef-CLIP identifies small objects accurately while D3Net has false detections of large objects (row 3,5).

References

[1] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D3Net: A unified speaker-listener architecture for 3D dense captioning and visual grounding. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2022.

[2] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. PointGroup: Dual-set point grouping for 3D instance segmentation. In *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2020.

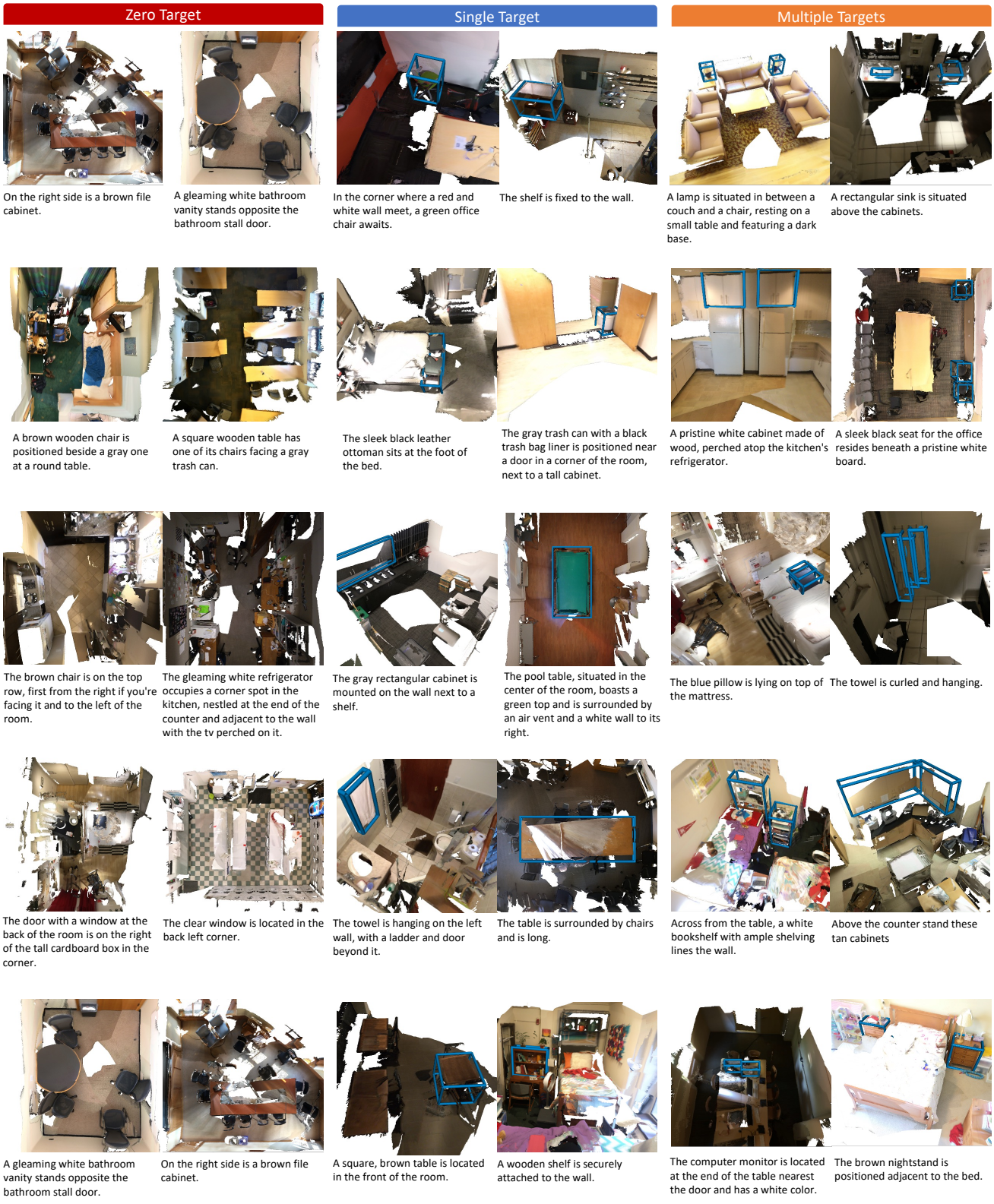

Figure 3: Examples of scene-description pairs with Zero Target, Single Target, and Multiple Targets from our Multi3DRefer dataset. Blue boxes indicate GT.

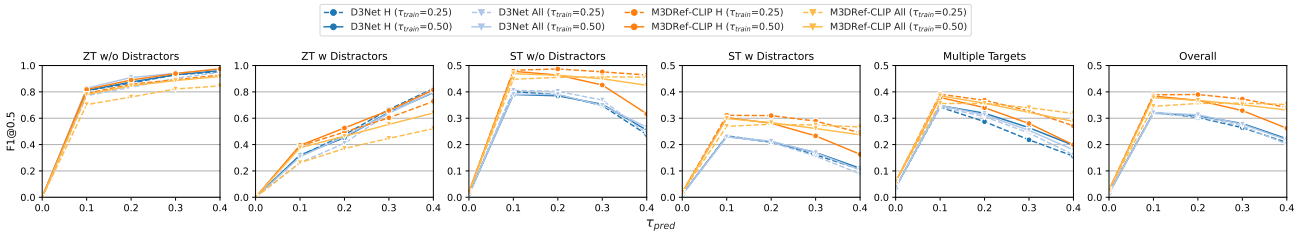

Figure 4: F1@0.50 on Multi3DRefer for the two methods with different matching strategies during training (All, Hungarian) and different values of τ_{pred} (x-axis), τ_{train} (solid=0.5, dashed=0.25). As we increase the prediction threshold τ_{pred}, we can get perfect performance on ZT cases (as nothing will ever be predicted). However, performance for ST and MT cases will drop. We find $\tau_{\text{pred}} = 0.1$ to be the optimal value.

Figure 5: Qualitative results of D3Net [1] versus M3DRef-CLIP on Multi3DRefer using predicted boxes. Blue boxes indicate GT, green boxes are true positives with IoU threshold $\tau_{\text{pred}} > 0.5$. Red boxes are false positives.