Appendix A. Implementation Details

Appendix A.1. Fine-tuning on downstream tasks

3D visual grounding. We adapt the pre-trained model for 3D visual grounding by replacing the cross-modal fusion decoder with a grounding head. Specifically, the grounding head consists of two layers of cross-attention and one self-attention layer between them. We adopt the AdamW [6] optimizer with cosine learning rate decay strategy. The initial learning rate is set to 5e-4 for the grounding head and 1e-4 for the rest parts. We follow [2] to train with the batch size of 10 and use 8 sentences for each input point cloud during training. The weight decay factor is 1e-5.

3D dense captioning. We use the same captioning head as [2] to generate each captioning word in a recurrent structure. We set the initial learning rate to 5e-4 for the captioning head and 1e-4 for other modules. And the remaining detailed implementations are same as [2].

3D question answering. Following [1], we adopt a transformer-based fusion layer followed by MLPs as the QA decoder. We use the AdamW optimizer with initial learning rate of 1e-4. The learning rate is decreased by 0.2 after 15 epochs. During training, the batch size is 16 and we use the same data augmentation (*i.e.*, random rotation and translation) as [1].

Appendix A.2. Details for raw-point reconstruction

In the main paper, we conduct experiments on reconstruction of raw points in the ablation study (Sec. 4.3). Specifically, we first utilize the center point of each proposal to query for 64 nearest points in the input point cloud, and then reconstruct the *xyz* or *xyz+RGB* of the nearest points using corresponding proposal feature. For *xyz*, we calculate the chamfer distance [5] for the predicted neighboring points of each proposal as the reconstruction loss. And for *RGB*, we use $\mathcal{L}_1$ loss during network training.

Appendix B. Ablation on loss weights

$\mathcal{L}_{\text{det}}$, $\mathcal{L}_{\text{lang}}$ and $\mathcal{L}_{\text{match}}$ are used to train 3D and language encoders, and we follow the weights in [2] to balance them. $\mathcal{L}_{\text{CSA}}$ and $\mathcal{L}_{\text{M3LM}}$ are pre-training losses to learn unified 3D-language representation, and we simply select weights to ensure that the initial loss values are within the same range. Tab. 1 shows that these loss weights are quite stable, and they are fixed for all our experiments.

Appendix C. Analysis of loss functions

For the whole pipeline, the pre-training loss consists of $\mathcal{L}_{\text{CSA}}$ and $\mathcal{L}_{\text{M3LM}}$, and they improved Acc@0.5 by 1.19 and 1.26 respectively (see paper Table 4), which indicates they contribute equally to the overall framework. In CSA, we observe from Tab. 2 that $\mathcal{L}_{\text{SA}}$ enables more improvement on grounding (+0.57 Acc@0.5) while $\mathcal{L}_{\text{CA}}$ improves captioning more (+0.85 C@0.5). In M3LM, the MPM is more crucial than MLM as it consistently achieves more improvements on three downstream tasks. This is because 3D point clouds are naturally unstructured and masked reasoning on visual input is more important than standard MLM. Although different losses have varying contribution to different metrics, they are all crucial to learn universal, generic, and transferable representations across variety of 3D vision-language tasks.

Table 1. Influence of different loss weights. A refers to Acc.

	A@0.25	A@0.5	C@0.5		A@0.25	A@0.5	C@0.5
$1\mathcal{L}_{\text{CSA}}$+$0.2\mathcal{L}_{\text{M3LM}}$	51.33	39.36	54.55	$5\mathcal{L}_{\text{CSA}}$+$0.1\mathcal{L}_{\text{M3LM}}$	51.32	39.24	53.69
$3\mathcal{L}_{\text{CSA}}$+$0.2\mathcal{L}_{\text{M3LM}}$	51.37	39.41	54.90	$5\mathcal{L}_{\text{CSA}}$+$0.2\mathcal{L}_{\text{M3LM}}$	**51.41**	39.46	**54.94**
$5\mathcal{L}_{\text{CSA}}$+$0.2\mathcal{L}_{\text{M3LM}}$	**51.41**	**39.46**	**54.94**	$5\mathcal{L}_{\text{CSA}}$+$0.5\mathcal{L}_{\text{M3LM}}$	51.37	**39.49**	54.50

Table 2. Influence of different loss functions.

Loss	Acc@0.25	Acc@0.5	C@0.5	EM@1
CSA	50.33	38.20	52.11	21.01
-CA	50.19 (⇓**0.14**)	38.10 (⇓0.10)	51.26 (⇓**0.85**)	20.86 (⇓0.15)
-SA	50.22 (⇓0.11)	37.63 (⇓**0.57**)	51.59 (⇓0.52)	20.78 (⇓**0.23**)
M3LM	51.41	39.46	54.94	21.65
-MPM	50.81 (⇓**0.60**)	38.91 (⇓**0.55**)	51.25 (⇓**3.69**)	21.28 (⇓0.37)
-MLM	51.27 (⇓0.14)	39.12 (⇓0.34)	53.59 (⇓1.35)	21.24 (⇓**0.41**)

Appendix D. More Qualitative Comparisons

Due to the length limitation of the main paper, we have only presented qualitative comparisons to show our method significantly surpasses training from scratch. We further provide qualitative comparisons with state-of-the-art methods [1, 2] on 3D visual grounding (Fig. 1), 3D dense captioning (Fig. 2) and 3D question answering (Fig. 3). In Fig. 1, our method achieves more accurate localization results than 3DJCG [2] on ScanRefer dataset [3], especially when the description text is long. The generated captions in Fig. 2 show our method describes target objects more correctly (see the underlined parts) in terms of the spatial relation with other objects in the scene. Fig. 3 indicates our method performs better than ScanQA model [1] in both localizing target bounding boxes according to questions (see the left 3 columns) and answering questions about attributes and relations (see the 4th and 5th columns).

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 1, 3

[2] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3DJCG: A unified framework for joint dense captioning and visual grounding on 3D point clouds. In *Proceedings of*

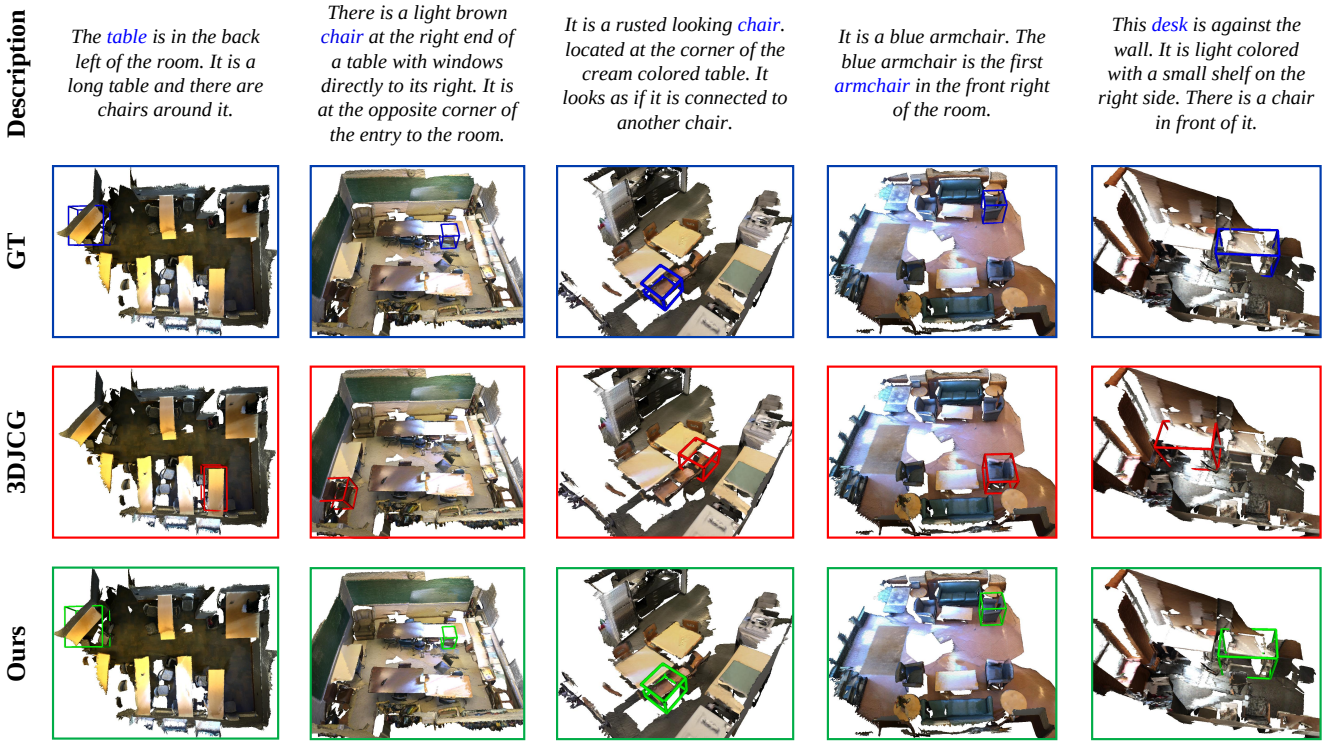

Figure 1. Qualitative comparison of 3D visual grounding results on ScanRefer [3] dataset. Blue, red and green represent the ground-truth (GT) label (*i.e.*, target bounding box), predicted results of 3DJCG [2] and ours, respectively.

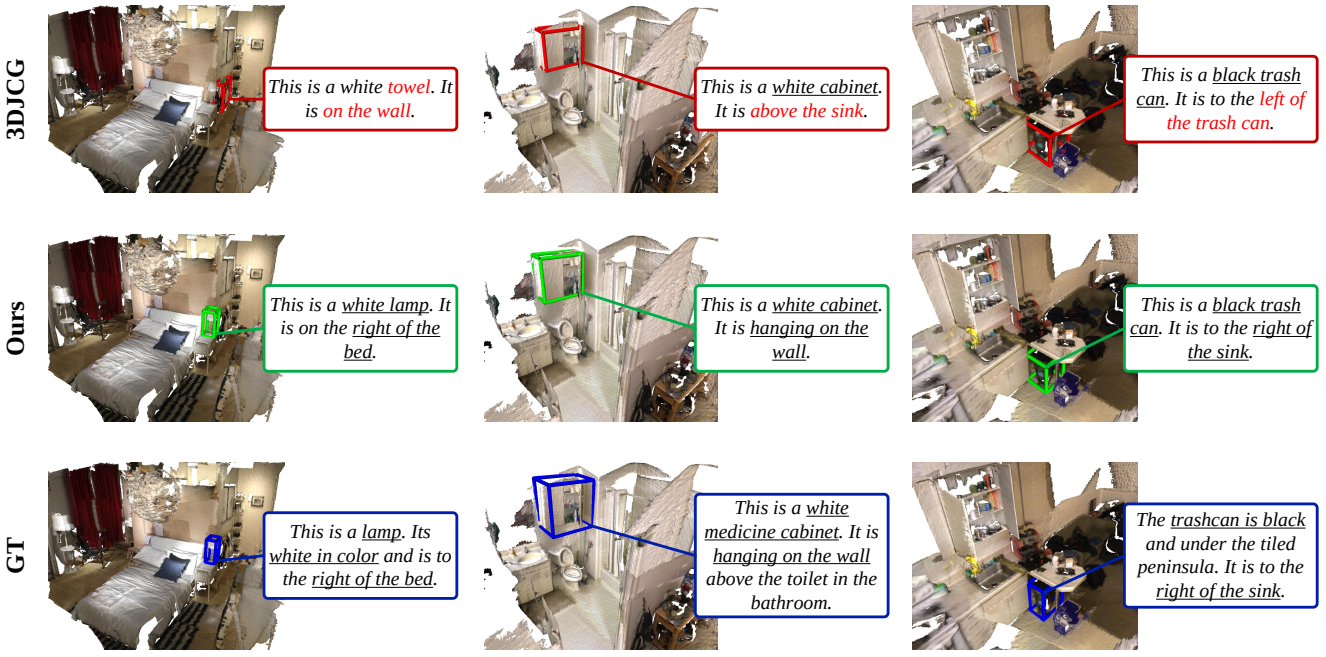

Figure 2. Qualitative comparison of 3D dense captioning results on Scan2Cap [4] dataset. The accurate parts of generated captions that matches ground-truth are underlined and the inaccurate parts are in red.

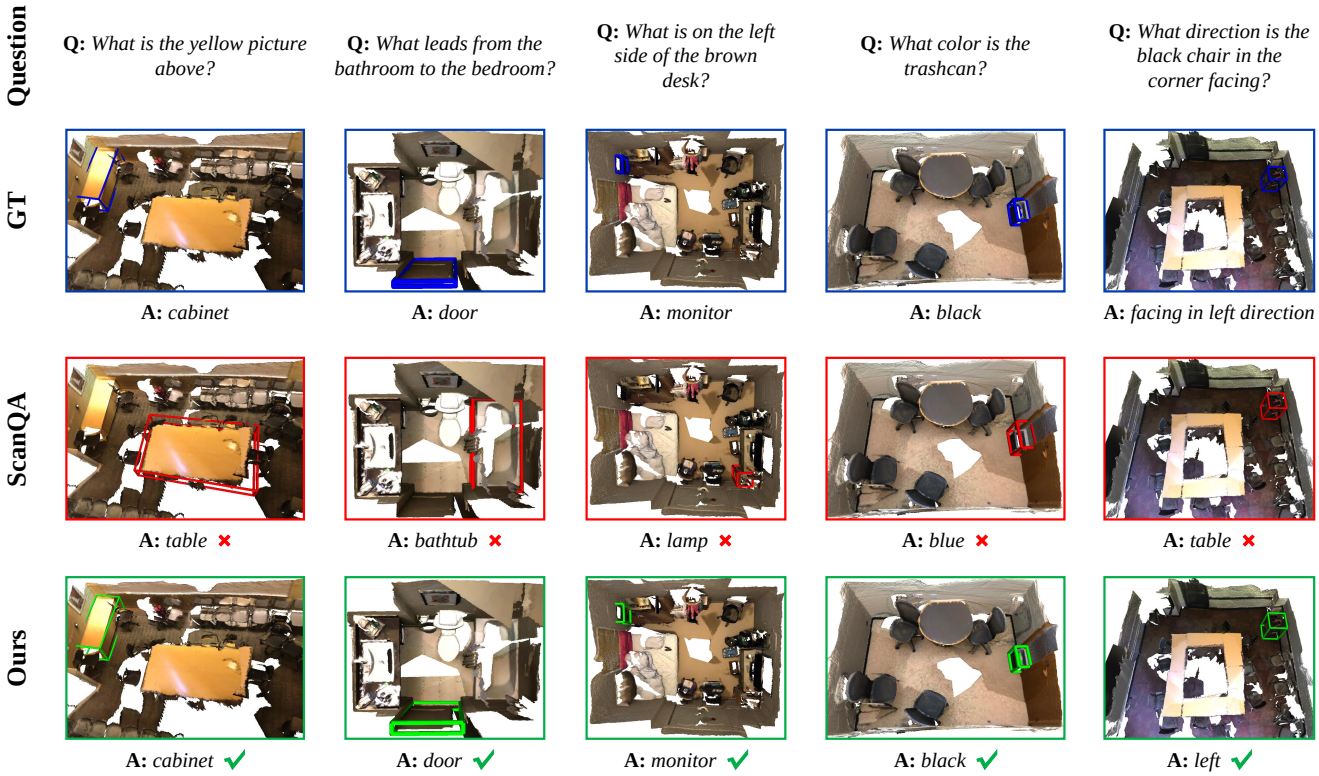

Figure 3. Qualitative comparison of 3D question answering results on ScanQA [1] dataset. Blue, red and green represent the ground-truth (GT) label (*i.e.*, answer text and bounding box), predicted results of ScanQA model [1] [2] and ours, respectively.

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16464–16473, 2022. 1, 2, 3

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *European Conference on Computer Vision (ECCV)*, pages 202–221, 2020. 1, 2

[4] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2021. 2

[5] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 605–613, 2017. 1

[6] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. In *International Conference on Learning Representations (ICLR)*, 2018. 1