Supplementary Material: CLIP-Guided Vision-Language Pre-training for Question Answering in 3D Scenes

Maria Parelli*, Alexandros Delitzas*, Nikolas Hars, Georgios Vlassis,
Sotirios Anagnostidis, Gregor Bachmann, Thomas Hofmann
ETH Zurich, Switzerland
{mparelli, adelitzas, nihars, gvlassis, sanagnos, gregorb}@ethz.ch

A. Details on the pre-training loss function

As discussed in Section 3.1.2, the combined loss used for pre-training the 3D Scene Encoder consists of three terms, i.e.,

$$\mathcal{L} = \mathcal{L}_{det} + \alpha\mathcal{L}_{text} + \beta\mathcal{L}_{image} \tag{1}$$

The first term of Eq. (1) comprises the object detection loss as introduced in [3] and is defined as

$$\begin{aligned}\mathcal{L}_{det} = \mathcal{L}_{vote\text{-}reg} + 0.5\mathcal{L}_{objn\text{-}cls} \\ + \mathcal{L}_{box} + 0.1\mathcal{L}_{sem\text{-}cls}\end{aligned} \tag{2}$$

where $\mathcal{L}_{vote\text{-}reg}$ represents the vote regression loss, $\mathcal{L}_{objn\text{-}cls}$ represents the objectness binary classification loss, $\mathcal{L}_{box}$ represents the box regression loss and $\mathcal{L}_{sem\text{-}cls}$ represents the semantic classification loss for the 18 ScanNet classes. For additional information, please refer to the respective publication [3].

The second and third terms of Eq. (1) can be defined as

$$\mathcal{L}_{text} = 1 - \cos\left(Z_{text}, Z_{scene}\right) \tag{3}$$

and

$$\mathcal{L}_{image} = 1 - \cos\left(Z_{image}, Z_{scene}\right) \tag{4}$$

respectively, where $\mathcal{L}_{text}$ is the cosine distance between Z_{text} and Z_{scene}, and $\mathcal{L}_{image}$ is the cosine distance between Z_{image} and Z_{scene}.

B. Experimental setup

In this section, we provide more details about the datasets and our experimental setup. In the pre-training stage, we utilize the ScanRefer [2] train set, which consists of 36,665 descriptions from 562 ScanNet scenes. In the training stage, we use the train set of ScanQA [1], which contains 25,563 questions from 562 ScanNet scenes. Note that both datasets follow the same train/test split as ScanNet. For question answering, we report our results on the two ScanQA test splits (with and without object annotations), which are hosted on EvalAI[1]. Since we do not have direct access to ground truth target object annotations for the test splits, we report our method's object localization performance on the ScanQA validation dataset.

References

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

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3D object localization in RGB-D scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer, 2020. 1

[3] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep Hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

*Equal contribution.

[1] https://eval.ai/web/challenges/challenge-page/1715/overview