Multi-CLIP: Contrastive Vision-Language Pre-training for Question Answering tasks in 3D Scenes (Supplementary Material)

Alexandros Delitzas*
adelitzas@ethz.ch
Maria Parelli*
mparelli@ethz.ch
Nikolas Hars
nihars@ethz.ch
Georgios Vlassis
gvlassis@ethz.ch
Sotirios Anagnostidis
sotirios.anagnostidis@inf.ethz.ch
Gregor Bachmann
gregor.bachmann@inf.ethz.ch
Thomas Hofmann
thomas.hofmann@inf.ethz.ch

ETH Zurich
Zurich, Switzerland

Overview. This supplementary material provides additional information and technical details. It is organized as follows: Section A provides additional explanations on the pre-training loss objective. Next, Section B and Section C contain additional experimental and implementation details respectively. Finally, in Section D, we further explain how we implemented the cosine similarity-based objective that was used in the ablation studies.

A Details on the pre-training loss objective

As discussed in Section 3.1 of the main paper, we utilize the following contrastive objective for pre-training the 3D scene encoder:

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

In Equation 1, the term $\mathcal{L}_{det}$ comprises the object detection loss, as introduced in [2], which can be defined as

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

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 seman-

*These authors contributed equally and are listed alphabetically.

tic classification loss for the 18 ScanNet classes. For additional information, we refer the interested reader to the corresponding work [4].

B Experimental setup

In this section, we provide more details about the datasets and our experimental setup.

B.1 Datasets

In the pre-training stage of the 3D scene encoder, we utilize only the train set of the ScanRefer [2] dataset, which consists of 36,665 descriptions from 562 ScanNet scenes. In the training stage of the 3D-VQA model, we use the train set of ScanQA [1], which contains 25,563 questions from 562 ScanNet scenes. In the training stage of the 3D-SQA model, we use the train set of SQA3D [3], which contains 16,229 situation descriptions alongside 26,623 questions from 518 ScanNet scenes. We note that all the aforementioned datasets follow the same train split as ScanNet. The train set of SQA3D contains a subset of the ScanNet train set since some scenes were purposely dropped during the dataset formation.

B.2 Evaluation

For the task of 3D-VQA, 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 access to ground truth target object annotations for the two test splits, we report our method's referred object localization performance on the ScanQA validation dataset.

For the task of 3D-SQA, we report our results on the test set of the SQA3D dataset, to which direct public access is provided by the authors. Since we have direct access to the test set of this dataset, we also use it to perform our ablation studies.

C Implementation details

The hyper-parameters used during the pre-training and training phases can be seen in Table 1 and Table 2 respectively. All experiments were conducted on a single NVIDIA GeForce RTX 3090 Ti GPU (24GB).

D Details on the ablation studies

In the Section 4.3 and 4.4 of the main paper, we compare our pre-training contrastive objective against a cosine similarity-based objective. To implement the latter objective, we alter the terms $\mathcal{L}_{text}$ and $\mathcal{L}_{image}$ of the proposed objective as following:

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

and

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

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

	Hyper-parameter	Value (pre-training)
Pre-training	Batch size	16
	Steps	15e3
	Weight decay	1e-5
	Initial learning rate	1e-4
	Gradient norm clipping	1
Adam	β_1	0.9
	β_2	0.999
	ε	1e-8
Loss	α	0.5
	β	0.5

Table 1: Hyper-parameters used in the pre-training stage.

	Hyper-parameter	Fine-tuning (3D-VQA)	Fine-tuning (3D-SQA)
Training	Batch size	16	16
	Epochs	40	50
	Weight decay	1e-5	1e-5
	Initial learning rate	5e-4	5e-4
	Learning rate decay	0.2	0.2
	Learning rate decay epoch	15	15
	Gradient norm clipping	1	1
Adam	β_1	0.9	0.9
	β_2	0.999	0.999
	ε	1e-8	1e-8
Model	M (object proposals)	256	256
	h (hidden dimension)	256	256

Table 2: Hyper-parameters used in the training stage.

where $\mathcal{L}'_{text}$ is the cosine distance between the text and 3D scene representations, i.e., Z_{text} and Z_{scene} respectively, and $\mathcal{L}'_{image}$ is the cosine distance between the 2D image and 3D scene representations, i.e., Z_{image} and Z_{scene} respectively. Therefore, the final cosine similarity-based loss can be formulated as

$$\mathcal{L}'_{pre} = \mathcal{L}_{det} + \alpha \mathcal{L}'_{text} + \beta \mathcal{L}'_{image} \tag{5}$$

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.

[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.

[3] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

[4] 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.