Towards Efficient Large-Scale Language-3D Representation Learning

Shentong Mo [1] Xiaogang Xu [2] Tongzhou Wang [3] Antonio Torralba [3] Shuang Li [3]

Abstract

Recent years have seen significant advancements in large-scale representation learning of 2D vision and language tasks. However, the efficacy of such cross-modal training on large-scale 3D objects with other modalities (such as text and images) remains primarily unexplored. We introduce MaskCL3D, an efficient and powerful method for language-3D representation learning. By employing contrastive and masked reconstruction learning on 3D point clouds with language descriptions and multi-view images, MaskCL3D significantly boosts training and testing efficiency. Meanwhile, we collect a large-scale language-3D dataset covering a wide array of objects and descriptions. Models trained on our dataset consistently outperform those trained on alternative datasets. Compared to existing baselines, our approach trained on the new dataset achieves state-of-the-art performance on zero-shot classification and retrieval tasks. We have performed a series of analytical studies on the learned language and 3D representations and find that these representations contain rich semantic information, which is crucial for interpreting and correlating intricate concepts within 3D environments.

1. Introduction

Language-3D representation learning (Huang et al., 2023; Hegde et al., 2023; Cheraghian et al., 2022; Xue et al., 2023a; Qi et al., 2023; Liu et al., 2023) is a rapidly emerging field that aims to bridge the gap between linguistic and spatial understanding. By integrating these two domains, this study seeksto enable more intuitive and effective interactions between humans and 3D environments, essential for applications in augmented reality, robotics, and virtual assistants. The ability to comprehend and manipulate 3D objects based on language commands is vital for creating more natural and user-friendly interfaces in these technologies.

However, the task of training robust 3D language representations is fraught with challenges. Chief among these is the computational intensity required to process 3D data. Training on detailed 3D models, such as meshes with a high number of vertices and faces, is not just resource-intensive but also time-consuming. This complexity significantly slows down the process of developing and fine-tuning models that can efficiently handle 3D language tasks. Another challenge is the scarcity of comprehensive 3D knowledge. Existing datasets and models often lack the necessary range and complexity for effectively modeling language-3D.

To address these obstacles, we introduce a simple yet effective approach, named MaskCL3D, for language-3D representation learning. To tackle the first challenge of computational demand and speed, MaskCL3D employs contrastive training and masked reconstruction, significantly enhancing training and testing efficiency. Our method allows for effective training with just a minimal subset of 3D points from the 3D mesh, leading to training faster than traditional methods. Remarkably, the proposed method achieves superior performance despite the reduced number of points used.

To tackle the challenge of scarce 3D knowledge, our method adopts a dual-encoder strategy: it incorporates pre-trained encoders for harnessing extensive commonsense knowledge and a newly crafted model encoder for assimilating specific, novel 3D knowledge. Large pre-trained language (LLM) and 2D image encoders are pivotal in retaining rich semantics, playing a crucial role in learning zero-shot representation. Therefore, we leverage these pre-trained encoders to facilitate the training of 3D representations, harnessing their capacity to enrich the semantic depth and accuracy of our 3D models. This approach not only improves the quality of 3D representations but also bridges the gap between language, 2D images, and 3D learning paradigms.

To overcome the scarcity of 3D data for large-scale language-3D representation learning, we have created the Caption-Objaverse dataset. This dataset comprises 700k carefully curated language-3D pairs, providing a rich and diverse resource for advanced study in this field. We demonstrate that methods trained on Caption-Objaverse outperform those trained on other datasets. This indicates the superior quality and effectiveness of Caption-Objaverse in enhancing model performance in language-3D representation learning. Our method, trained on Caption-Objaverse, markedly surpasses state-of-the-art methods, achieving a notable 7.69% improvement over baselines in zero-shot recognition.

[1] Carnegie Mellon University [2] Zhejiang University [3] Massachusetts Institute of Technology. Correspondence to: Shentong Mo <shentongmo@gmail.com>.

Proceedings of the 1^{st} Workshop on Efficient Systems for Foundation Models, Vienna, Austria. 2024. Copyright 2024 by the author(s).

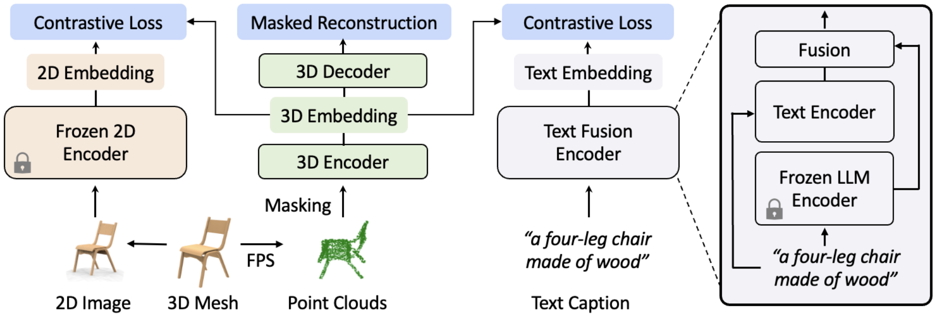


Figure 1: **Pipeline of the proposed method**. To enable cross-domain representation learning between 3D point clouds and language descriptions, as well as improve training and evaluation efficiency, we employ both contrastive training and masked reconstruction training in our framework. This mechanism involves sampling a subset of 3D points for joint feature learning using contrastive loss between 3D-2D and 3D-Text embeddings while using the remaining points for 3D reconstruction with masked reconstruction loss.

To gain deeper insights into the representations learned by our model, we carried out analysis experiments aimed at evaluating the effectiveness of these learned representations. Remarkably, our analysis reveals that these representations possess the ability for arithmetic and compositional operations. For instance, the 3D-text similarity between "a yellow bandage – yellow + red" closely aligns with that of "a red bandage". This capability highlights the advanced semantic understanding and flexibility of our model in interpreting and relating complex concepts in 3D environments.

2. Method

Given the point cloud of a 3D object and the caption of the same object, we learn joint 3D and language representations. We first use the 3D object encoder to encode the 3D point clouds into a feature representation. Then we adopt the frozen 2D vision encoder pre-trained from CLIP (Radford et al., 2021) to extract 2D embeddings from multi-view images of the 3D object. Similarly, we encode the language description into a feature representation through a text fusion encoder. We adopt the contrastive loss (Radford et al., 2021) to train the model (Sec. 2.1). This loss function facilitates the convergence of positive language-3D pairs, bringing them closer together, while simultaneously pushing negative feature pairs further apart. Consequently, our model learns to capture the intrinsic relationships between the 3D data and the corresponding language descriptions.

Training a model on 3D point clouds with a large-scale dataset size can be computationally expensive, particularly as the number of points increases. To address this challenge and improve both training and evaluation efficiency, we propose a masked reconstruction loss on 3D point clouds (Sec. 2.2). Remarkably, our proposed method with the masked reconstruction loss not only achieves superior performance compared to the model without this enhancement but also significantly accelerates the training process.

Overall, our approach presents a comprehensive framework for learning joint 3D and textual representations from large-scale data, leveraging 3D point cloud data and captions. By combining contrastive and masked reconstruction loss, we achieve efficient training, improved performance, and accelerated evaluation, advancing zero-shot tasks.

2.1. Contrastive Training

3D Encoder. Let $\mathcal{D} = \{(t_i, X_i) : i = 1, ..., N\}$ be a dataset of texts and 3D point clouds. Each point cloud is sampled from the object 3D mesh and consists of p points, denoted as $X_i \in \mathbb{R}^{p\times 3}$, where the three dimensions represent the XYZ coordinates of each point. Each point cloud is sent to a Transformer model (Vaswani et al., 2017b) to encode the 3D information. We take the output of the last Transformer layer to represent the point cloud, denoted as $f_{X_i} \in \mathbb{R}^d$.

Text Fusion Encoder. To extract prior knowledge about the textual description of the object, we first tokenize the words in the sentence and then pass them through a pre-trained LLM encoder, which provides an initial textual feature representation denoted as $g'_{t_i} \in \mathbb{R}^d$. To further capture the new 3D text knowledge on the training data, we introduce a new text encoder, which is trained from scratch. This text encoder encodes the textual information into a feature vector denoted as $g''_{t_i} \in \mathbb{R}^d$. By combining the pre-trained textual feature with the newly learned feature, we obtain the final textual representation $g_{t_i} = g'_{t_i} + g''_{t_i}$. This approach allows us to incorporate both the commonsense information from the pre-trained language model and the novel 3D information derived from a large amount of 3D data.

Image Encoder. For multi-view 2D images, we adopt a frozen 2D encoder from the vision encoder (ViT-L) in CLIP (Radford et al., 2021) to generate averaged 2D embeddings from the last Transformer layer, denoted as $f_{I_i} \in \mathbb{R}^d$.

Constrative Training Loss. To facilitate cross-domain feature representation learning, we employ a contrastive loss to align 3D point clouds with textual descriptions and images. The contrastive loss is defined as follows:

$$\begin{aligned}\mathcal{L}_C = \frac{1}{N}\sum_{i=1}^{N} &-\log \frac{\exp\left(\frac{1}{\tau}s(f_{X_i}, g_{t_i})\right)}{\sum_{j=1}^{N}\exp\left(\frac{1}{\tau}s(f_{X_j}, g_{t_i})\right)} \\ &-\log \frac{\exp\left(\frac{1}{\tau}s(f_{X_i}, f_{I_i})\right)}{\sum_{j=1}^{N}\exp\left(\frac{1}{\tau}s(f_{X_j}, f_{I_i})\right)},\end{aligned} \tag{1}$$

where s represents the cosine similarity between two feature

Table 1: **Zero-shot recognition results for models trained on ShapeNet (Chang et al., 2015).** Our method, utilizing fewer training and testing points, is faster than the baseline methods and surpasses their performance on the same training dataset.

Method	**Pretrain Data**	**Pretrain Modalities**	**# Train Points**	**# Test Points**	**ModelNet10**	**ModelNet40**
PointCLIP (Zhang et al., 2022)[CVPR 2022]	ShapeNet	Text + 2D	–	–	59.30	19.30
CLIP2Point (Huang et al., 2023)[ICCV 2023]	ShapeNet	Text + 2D	–	–	66.63	49.38
PointCLIP v2 (Zhu et al., 2023)[ICCV 2023]	ShapeNet	Text + 2D	–	–	73.13	64.22
CG3D (Hegde et al., 2023)[arXiv 2023]	ShapeNet	Text + 2D + 3D	8192	8192	67.30	50.60
Cheraghian *et al.* (Cheraghian et al., 2022)[IJCV'2022]	ShapeNet	Text + 2D + 3D	4096	4096	68.50	–
ULIP (Xue et al., 2023a)[CVPR 2023]	ShapeNet	Text + 2D + 3D	2048	2048	73.87	60.40
ReCon (Qi et al., 2023)[ICML 2023]	ShapeNet	Text + 2D + 3D	1024	8192	75.60	61.70
ULIP-2 (Xue et al., 2023b)[arXiv 2023]	ShapeNet	Text + 2D + 3D	2048	2048	–	66.40
OpenShape (Liu et al., 2023)[NeurIPS 2023]	ShapeNet	Text + 2D + 3D	4096	4096	–	70.30
MaskCL3D (ours)	ShapeNet	Text + 2D + 3D	**512**	**1024**	**81.15**	**71.08**

Table 2: **Zero-shot recognition results for models trained on Caption-Objaverse.** Our method largely outperforms all the baselines.

Method	**Pretrain Data**	**Pretrain Modalities**	**# Train Points**	**# Test Points**	**ModelNet10**	**ModelNet40**
ULIP (Xue et al., 2023a)[CVPR 2023]	Caption-Objaverse	Text + 3D	2048	2048	64.53	51.82
MaskCL3D (ours)	Caption-Objaverse	Text + 3D	**512**	**1024**	**77.26**	**65.78**
ULIP (Xue et al., 2023a)[CVPR 2023]	Caption-Objaverse	Text + 3D + 2D	2048	2048	76.39	64.27
ReCon (Qi et al., 2023)[ICML 2023]	Caption-Objaverse	Text + 3D + 2D	1024	8192	78.21	65.85
OpenShape (Liu et al., 2023)[NeurIPS 2023]	Caption-Objaverse	Text + 3D + 2D	4096	4096	80.26	67.93
ULIP-2 (Xue et al., 2023b)[arXiv 2023]	Caption-Objaverse	Text + 2D + 3D	2048	2048	–	74.16
MaskCL3D (ours)	Caption-Objaverse	Text + 3D + 2D	**512**	**1024**	**85.35**	**75.62**

vectors, while τ denotes the temperature hyperparameter. The contrastive loss serves the purpose of bringing positive pairs (consisting of matching 3D point clouds f_{X_i} and textual descriptions g_{t_i} and matching 3D point clouds f_{X_i} and 2D images f_{I_i}) closer together while pushing negative pairs (comprising non-matching instances $[f_{X_j}, g_{t_i}]$ and $[f_{X_j}, f_{I_i}]$, where $j \neq i$) further apart. This encourages the model to capture the underlying relationships and associations between the 3D structures and the corresponding language description and multi-view 2D images.

2.2. Masked Reconstruction Training

Training a model on a large set of 3D point clouds can be computationally expensive, particularly as the number of points increases. We thus propose a masked reconstruction loss to improve the training and evaluation efficiency. Following (Pang et al., 2022), we apply Farthest Point Sampling (FPS) (Eldar et al., 1997) to sample point clouds from 3D mesh. FPS selects c samples as centers and applies the K-Nearest Neighborhood (KNN) algorithm to select k nearest points for each center, resulting in an initial point cloud represented as $X_i \in \mathbb{R}^{c \times k \times 3}$. It is important to note that each point in the point cloud is normalized relative to its corresponding center point. This normalization step facilitates better convergence during the training phase.

For efficient training and evaluation, we introduce a masking ratio denoted as r, which randomly removes a portion of points from each point cloud cluster during training. The remaining visible points are then forwarded to the 3D encoder to generate a 3D embedding of the object as shown in Figure 1. A lightweight decoder Transformer is then trained to reconstruct the missing 3D points. This decoder is designed to be smaller than the encoder, enabling more efficient training. The masked reconstruction loss, denoted as $\mathcal{L}_M$, is defined as follows:

$$\begin{aligned} \mathcal{L}_M = & \frac{1}{|\widehat{X}_i^m|} \sum_{\hat{x} \in \widehat{X}_i^m} \min_{x \in X_i^m} \|\hat{x} - x\|_2^2 \\ & + \frac{1}{|X_i^m|} \sum_{x \in X_i^m} \min_{\hat{x} \in \widehat{X}_i^m} \|\hat{x} - x\|_2^2, \end{aligned} \tag{2}$$

where X_i^m denotes the set of removed points and $\widehat{X}_i^m$ denotes the corresponding prediction of the decoder model. To evaluate the accuracy of the point cloud reconstruction, we compute the distance between the ground truth XYZ coordinates of a point x in X_i^m and its corresponding predicted XYZ coordinates $\hat{x}$. This distance measurement allows us to quantify the dissimilarity between the original and reconstructed points, providing insight into the quality of the masked reconstruction process.

The proposed masked reconstruction loss allows us to reduce the number of points used for both training and testing. When compared with the current state-of-the-art model, OpenShape (Liu et al., 2023), our approach utilizes 87.5% fewer points during training and 75% fewer points during inference phases. Remarkably, despite this substantial reduction in data points, MaskCL3D consistently surpasses OpenShape's performance, achieving an impressive 5% to 9% improvement across various datasets. This approach effectively accelerates the training and testing process, allowing for more efficient model convergence while maintaining the ability to handle missing information in the point clouds.

2.3. Joint Training

To achieve the goals of cross-domain learning and efficient training simultaneously, we jointly optimize both objectives:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_M. \tag{3}$$

By optimizing the combined loss $\mathcal{L}$, the model can simultaneously learn to generate informative and discriminative cross-domain feature representations and effectively reconstruct missing points in the point clouds.

3. Experimental Evaluations

3D Recognition. We evaluate the accuracy of predicted text labels of 3D point clouds. It is important to note that there is no overlap between the training and testing objects. This setup allows us to evaluate the generalization capability of each approach when encountering unseen objects. We first show the results of models trained on the commonly used 3D-language dataset, ShapeNet, in Table 1. The results on the ModelNet10 and ModelNet40 datasets are reported. Our MaskCL3D employs only 512 points during training

Table 3: **Text-to-3D and 3D-to-Text retrieval results on Caption-Objaverse.** Our method dramatically outperforms baselines.

Method	Text-to-3D			3D-to-Text		
	R@1 (↑)	**R@5(↑)**	**R@10(↑)**	**R@1 (↑)**	**R@5(↑)**	**R@10(↑)**
Point-CLIP (Zhang et al., 2022)[CVPR 2022]	15.6	34.7	39.8	15.8	35.1	40.6
ULIP (Xue et al., 2023a)[CVPR 2023]	25.5	46.3	57.8	25.9	46.9	58.5
ReCon (Qi et al., 2023)[ICML 2023]	27.8	47.5	58.6	28.5	48.2	59.3
OpenShape (Liu et al., 2023)[NeurIPS 2023]	30.6	49.2	60.3	31.2	49.5	60.9
MaskCL3D (ours)	**39.7**	**57.9**	**69.5**	**40.6**	**58.7**	**71.6**

Table 4: **Impact of varying numbers of test points**. MaskCL3D excels over baselines with fewer test points, demonstrating its efficiency and effectiveness.

# Test Points	256	1024	4096
Point-CLIP (Zhang et al., 2022)[CVPR 2022]	0.15	0.27	0.36
ULIP (Xue et al., 2023a)[CVPR 2023]	0.42	0.48	0.62
ReCon (Qi et al., 2023)[ICML 2023]	0.46	0.51	0.69
OpenShape (Liu et al., 2023)[NeurIPS 2023]	0.52	0.63	0.75
MaskCL3D (ours)	**0.69**	**0.78**	**0.86**

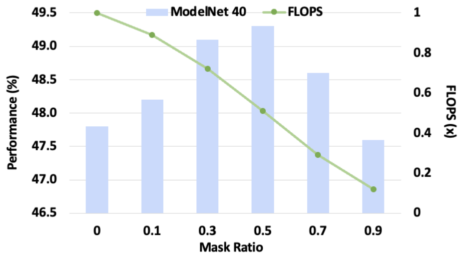


Figure 2: **Effect of Mask Ratio on Zero-Shot Recognition Accuracy and FLOPs on the ModelNet40.** Our default masking ratio of 0.5 leads to optimal performance while also reducing computational requirements, as evidenced by lower FLOPs.

and 1024 points during testing, which is significantly faster than the baseline methods. Additionally, our method shows improvements over these baselines.

We further compare MaskCL3D and baselines trained on our Caption-Objaverse dataset, as shown in Table 2. We observe that the same method, when trained on our dataset, exhibits higher performance compared to its training on the ShapeNet dataset in Table 1. This is because Caption-Objaverse contains a diversity of 3D objects and rich textual descriptions. Furthermore, employing a pre-trained 2D image encoder significantly enhances zero-shot performance due to its rich knowledge base. Our method significantly surpasses the baselines, exemplified by its 5.09% improvement over OpenShape on the ModelNet10 dataset. This demonstrates the effectiveness in zero-shot recognition tasks.

Text-3D Retrieval. The learned text and 3D feature representations can be used for Text-to-3D and 3D-to-Text retrieval. In Table 3, we evaluate the retrieval results of different methods on our Caption-Objaverse dataset. MaskCL3D achieves the best performance in all metrics compared to the baselines. In particular, MaskCL3D outperforms Point-CLIP (Zhang et al., 2022), the strong baseline using multi-view 2D depth maps, by 24.1 on R@1. Moreover, we obtain the performance gains of 14.2 on R@1, compared to ULIP (Xue et al., 2023a), the state-of-the-art work on 3D representation learning involving 2D images.

Impact of Varying Numbers of Test Points. While MaskCL3D model is trained with 1024 input points, it can readily be applied to more or fewer points in testing. We compare MaskCL3D and baselines using different numbers of points during testing. We randomly sample 500 pairs of 3D point clouds and language descriptions. We then compute the cosine similarity of the feature representations of the paired 3D point cloud and the corresponding language description. As demonstrated in Table 4, utilizing a greater number of test points can enhance performance, but it also results in longer inference times. MaskCL3D, even when using only 256 test points, already exceeds the performance of all baseline methods that use 1024 test points and outperforms the majority of baselines utilizing 4096 test points.

Masking Ratio & FLOPs. We explore the effects of varying mask ratios in the masked reconstruction training. MaskCL3D is trained on the Caption-Objaverse dataset and assessed on the ModelNet40 for zero-shot recognition tasks. This is different from the experiment in Section B.4, which examines the impact of different numbers of test points. In Figure 2, we find that using a ratio of 0.5 (512 points) achieves the best performance. As the masking ratio increases from 0 to 0.5, we observe a consistent improvement in the results. This is because masking a certain proportion of input yields a nontrivial and meaningful self-supervisory task for reconstructing masked parts , as demonstrated in MAE (He et al., 2021). However, a higher mask ratio leads to reduced performance, suggesting that maintaining an adequate number of points is essential for effective 3D representation learning. The FLOPs decrease as the mask ratio increases, since fewer points are utilized in the training process. Our method employs a masking ratio of 0.5, which not only achieves the best performance but also results in lower computational demands, as measured in FLOPs.

4. Conclusion

We present MaskCL3D, a simple yet effective approach for large-scale language-3D representation learning. It employs a blend of contrastive learning and masked reconstruction on 3D point clouds, captions, and multi-view images, significantly improving training and testing efficiency. We have also curated a comprehensive language-3D dataset, which significantly boosts the performance of our model.

Broader Impact. Our work thoroughly analyzes many capabilities of large-scale pre-training of language-3D representations from MaskCL3D. We believe that these are promising for diverse researchers in the 3D community.

References

Bao, H., Dong, L., Piao, S., and Wei, F. BEit: BERT pre-training of image transformers. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 10

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 10

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3, 8

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020a. 10

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020b. 10

Chen, X. and He, K. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c. 10

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 10

Cheraghian, A., Rahman, S., Chowdhury, T. F., Campbell, D., and Petersson, L. Zero-shot learning on 3d point cloud objects and beyond. *International Journal of Computer Vision*, 130:2364–2384, 2022. 1, 3, 8

Conneau, A. and Lample, G. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 10

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 7

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 9, 10, 13

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., and Yu, N. Peco: Perceptual codebook for BERT pre-training of vision transformers. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2023. 10

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 7, 11

Eldar, Y., Lindenbaum, M., Porat, M., and Zeevi, Y. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. doi: 10.1109/83.623193. 3

Feichtenhofer, C., Fan, H., Li, Y., and He, K. Masked autoencoders as spatiotemporal learners. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2022. 10

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 10

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. 10

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 4, 10

Hegde, D., Valanarasu, J. M. J., and Patel, V. M. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023. 1, 3, 8

Huang, T., Dong, B., Yang, Y., Huang, X., Lau, R. W., Ouyang, W., and Zuo, W. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. pp. 22157–22167, 2023. 1, 3, 8

Li, J., Zhou, P., Xiong, C., and Hoi, S. Prototypical contrastive learning of unsupervised representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 10

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a. 7

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b. 10

Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., and Su, H. Openshape: Scaling up 3d shape representation towards open-world understanding. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3, 4, 8, 10, 11

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 10

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 13

Luo, T., Rockwell, C., Lee, H., and Johnson, J. Scalable 3d captioning with pretrained models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 10

Mo, S., Sun, Z., and Li, C. Siamese prototypical contrastive learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. 10

Mo, S., Sun, Z., and Li, C. Rethinking prototypical contrastive learning through alignment, uniformity and correlation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 10

Pang, Y., Wang, W., Tay, F. E. H., Liu, W., Tian, Y., and Yuan, L. Masked autoencoders for point cloud self-supervised learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 3, 11, 13

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8026–8037, 2019. 13

Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., and Yi, L. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning (ICML)*, 2023. 1, 3, 4, 8, 10, 11

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 7, 10, 13

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 10

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 10

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588–1597, 2019. 8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017a. 10

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b. 2

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020. 9, 10

Wang, X., Liu, Z., and Yu, S. X. CLD: unsupervised feature learning by cross-level instance-group discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 10

Wei, C., Fan, H., Xie, S., Wu, C., Yuille, A. L., and Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 10

Wettig, A., Gao, T., Zhong, Z., and Chen, D. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2985–3000, 2023. 10

Wu, J. and Mo, S. Object-wise masked autoencoders for fast pre-training. *arXiv preprint arXiv:2205.14338*, 2022. 10

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912–1920, 2015. 8

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 10

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a. 1, 3, 4, 8, 10, 11, 13

Xue, L., Yu, N., Zhang, S., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. Ulip-2: Towards scalable multimodal pre-training for 3d understanding, 2023b. 3

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8, 11

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., and Li, H. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 8, 10

Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., and Gao, P. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2639–2650, 2023. 3, 8

Appendix

In this appendix, we provide

- Details about the Caption-Objaverse Dataset in Appendix A,
- Experimental evaluations in Appendix B,
- Additional analyses of the learned language and 3D representations in Appendix C,
- Discussions on realted work in Appendix D,
- More ablation studies and experimental results of the proposed method in Appendix E,
- Analyses of the proposed Caption-Objaverse dataset as shown in Appendix F,
- Implementation details in Appendix G.

A. The Caption-Objaverse Dataset

To address the limited training data for large-scale language-3D representation learning, we build the Caption-Objaverse dataset. Caption-Objaverse is based on Objaverse (Deitke et al., 2022), the largest 3D dataset with 700k publicly-released 3D objects. For ethical concerns, we removed around 19k objects with identifiable facial scans and NSFW content. We augment this dataset with language descriptions for each object. To do this, we project the 3D mesh of each object into the 2D image domain, and use the state-of-the-art image caption approach, BLIP-2 (Li et al., 2023a), to generate a caption for the 2D image. After generating the captions, we filtered out low-quality ones and updated the 3D object with new captions, resulting in $700,862$ text-3D pairs. Specifically, we calculated the CLIP (Radford et al., 2021) score between generated captions and 2D-rendered images and manually relabeled samples with a similarity score$<$0.5.

To future examine the generated caption, we add a text-to-image retrieval experiment using a different image encoder, ViT-B-32 (Dosovitskiy et al., 2021). The recall@10 is 92.65%, indicating the correct images can be effectively retrieved using the generated captions. To evaluate the dataset complexity, we built a test set of 500 samples with long captions (average length of 100 words). Our method trained on the original dataset performs well on the long-caption test set, achieving 81.63% on R@10 for text-to-image retrieval. This means the original dataset is not simple and models trained on it can understand complex instructions. We provide more details about datasets in the supplementary material.

B. Experimental Evaluations

B.1. Evaluation Datasets

We evaluate MaskCL3D and baselines on seven datasets. **ModelNet40** (Wu et al., 2015) is a synthetic 3D CAD benchmark with 40 object categories that include 9,843 3D models for training and 2,468 for testing. **ModelNet10** (Wu et al., 2015) is a subset of ModelNet40, containing 10 object categories. **ScanObjectNN** (Uy et al., 2019) has 2,902 scanned 3D objects from the real world, covering 15 categories. This benchmark has three variants for evaluation: **ObjectOnly** includes ground truth segmented objects extracted from the scene meshes; **ObjectBg** contains objects with background; **Hardest** includes objects with perturbations such as scaling, rotation, and translation. **ShapeNet** (Chang et al., 2015) consists of 50,000 distinct 3D objects from 55 common object categories. We use two subsets provided in (Yu et al., 2022): **ShapeNet34** with 34 classes and **ShapeNet55** with all 55 classes.

B.2. Evaluation Metrics

Accuracy of 3D recognition. Following previous work (Zhang et al., 2022; Xue et al., 2023a), we evaluate the zero-shot recognition performance of each method by computing distances between the 3D features of a given 3D object and a set of text descriptions. The text description with the smallest distance is used as the predicted class. The test example is regarded as correct if the selected text description corresponds to the ground truth class. We report the average accuracy of all test examples in each dataset.

Recall of Text-3D retrieval. For text-to-3D retrieval, we are given input captions to retrieve 3D objects from our Caption-Objaverse by computing the similarity between 3D embeddings and text embeddings. For 3D-to-text retrieval, we are given 3D objects to retrieve text captions from our Caption-Objaverse by computing the similarity between text embeddings and 3D embeddings. We compute the Recall at rank r (R@r) to measure the percentage of labels retrieved within the top r ranked predictions, where $r = 1, 5, 10$.

B.3. Baselines

We compare the proposed framework, MaskCL3D, with the state-of-the-art approaches for 3D-language representation learning. PointCLIP (Zhang et al., 2022) and PointCLIP v2 (Zhu et al., 2023) learn 3D representations by aligning 3D objects with 3D class texts and multi-view 2D depth maps that are manually generated from 3D point clouds. (Huang et al., 2023; Hegde et al., 2023; Cheraghian et al., 2022; Xue et al., 2023a; Qi et al., 2023; Liu et al., 2023) learned based on object triplets with 3D point clouds, text, and 3D projected images, where they rendered 60 multi-view mages from each CAD model for 3D representation learning. For each method, we test it on three different random seeds and report the averaged results under each evaluation metric.

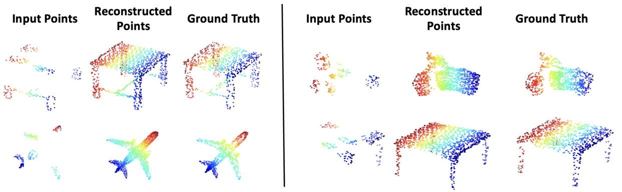


Figure 3: **Point clouds Reconstruction.** Our method can reconstruct the object points accurately even when 70% of the points are missing.

B.4. Zero-shot Results

Point Cloud Reconstruction. The proposed masked reconstruction training enables the reconstruction of object point clouds from incomplete object points, as illustrated in Figure 3. Our method can still reconstruct the object accurately even when 70% of the points are missing. These visualizations emphasize MaskCL3D's ability to accurately represent 3D information, even under conditions of partial data absence. We also evaluated the ℓ_2 Chamfer Distance ($\times 10^{-3}$) on ShapeNet for point cloud completion. Our method achieves lower reconstruction errors (2.832) than the strong baseline, ReCon (Qi et al., 2023) (5.785).

B.5. Ablation Studies

In this section, we examine the effectiveness of the proposed contrastive learning and masked reconstruction. Additionally, we perform ablation studies on the text fusion encoder and explore the influence of different masking ratios on zero-shot recognition tasks.

Contrastive Training & Masked Reconstruction Training. In Table 5, we examine the extent to which contrastive training and masked reconstruction training impact the results of zero-shot recognition tasks. We can observe that incorporating 3D-language contrastive training highly increases the results of zero-shot classification by 27.6% and 38.6% on ModelNet40 and ModelNet10, respectively. We also assess various masking strategies in masked reconstruction training. The term "Random" refers to the random sampling of object points during training. This strategy might inadvertently eliminate points crucial for 3D understanding, potentially impairing performance. "Learnable" has an embedding layer that is optimized with gradient descent across to select the most important points during training. The masking method we employ, Farthest Point Sampling (FPS), is designed to eliminate redundant points while preserving sufficient information for 3D understanding, thereby achieving optimal performance. These results confirm the importance of combining contrastive training and masked reconstruction training for language-3D representation learning.

Text Fusion Encoder. To utilize the pre-existing common-

Table 5: **Ablation studies of contrastive training (CL) and masked reconstruction training (Mask)**. Both CL and Mask are important in zero-shot recognition tasks. Using the FPS sampling strategy in the masked reconstruction training achieves the best performance.

CL	Mask	Type	ModelNet 40	ModelNet 10	ScanObjectNN-ObjectOnly	ScanObjectNN-ObjectBg	ScanObjectNN-Hardest	ShapeNet 34	ShapeNet 55
✗	✗	–	20.2	30.2	15.4	12.6	7.6	18.7	6.3
✓	✗	–	47.8	68.8	25.7	23.4	16.1	51.0	37.8
✓	✓	random	28.9	43.6	7.8	6.3	2.6	31.2	15.7
✓	✓	learnable	37.5	57.9	18.6	15.2	7.1	43.3	23.5
✓	✓	**FPS**	**49.3**	**70.9**	**27.3**	**26.2**	**17.5**	**53.7**	**39.2**

Table 6: **Ablation studies of the text fusion encoder.** Fusing the pre-trained LLM with the learned text encoder results in the best performance. "FT" represents a fine-tuned model, while "FZ" indicates a frozen model.

Pre-trained LLM	Text Encoder	Fusion	ModelNet 40	ModelNet 10	ScanObjectNN-ObjectOnly	ScanObjectNN-ObjectBg	ScanObjectNN-Hardest	ShapeNet 34	ShapeNet 55
✗	✓	–	47.8	68.5	25.8	23.1	16.3	51.2	37.5
FT	✗	–	48.1	68.9	26.1	24.2	16.9	52.3	38.3
FZ	✗	–	46.8	67.2	24.3	22.6	15.2	49.5	36.7
FZ	FT	cont	48.5	69.5	26.7	24.7	17.1	52.9	38.7
FZ	FT	seq	47.2	67.8	24.8	22.9	15.6	50.3	37.1
FZ	**FT**	**para**	**49.3**	**70.9**	**27.3**	**26.2**	**17.5**	**53.7**	**39.2**

sense knowledge from large language models (LLM) and integrate specific 3D information, our text fusion encoder combines the pre-trained language model, BERT (Devlin et al., 2018) base model, with a newly learned text encoder. In Table 6, we first assess the effectiveness of the pre-trained Large Language Model (LLM) by solely utilizing the learned text encoder for language embedding. The performance declines upon removing the pre-trained LLM. Next, we remove the learned text encoder and test both the fine-tuned (FT) and the frozen (FZ) versions of the LLM for language embeddings. The performance of both versions declines compared to the text fusion encoder employed in our method, as shown in the last row. We further evaluate the impact of different fusion types. In "concatenation (cont)", the text features from the pre-trained LLM and learned text encoder are concatenated along the dimension. In "sequential (seq)", the text is first encoded by the pre-trained LLM. The output of the pre-trained LLM is then sent to the learned text encoder. The "parallel (para)" refers to the process where the text feature from the pre-trained LLM is combined with the output from the learned text encoder, as illustrated in Figure 1. Merging the pre-trained LLM and the learned text encoder in a "parallel (para)" manner effectively combines pre-existing knowledge with 3D-specific knowledge.

C. Analyses of Learned Representations

We analyze the MaskCL3D language and 3D representations in terms of distributional properties, feature arithmetic, and compositionally.

Feature Distributions. Following (Wang & Isola, 2020), we use the Alignment and Uniformity metrics to evaluate the learned representations. The Alignment metric measures how well paired 3D and text features are aligned, while the Uniformity metric evaluates how well the representations of different classes are spread out. Table 7 compares our method against baselines. **Alignment** is measured as the squared ℓ_2-distance between features from text-3D positive pairs, where lower value indicates more aligned features[1]. MaskCL3D 3D representations are best aligned with text features among all methods. **Uniformity** is measured across randomly sampled (negative) features. MaskCL3D achieves the best uniformity metric, and thus better captures input 3D information compared to the baselines.

Feature Arithmetic. Our learned representations capture a wealth of semantic information. We find that these features support arithmetic operations, such as addition and subtraction. Take the snowman in Table 8 as an example. 3D-Text similarity between the snowman 3d model and text is merely 0.15 due to the incorrect information on the "bandage" in text. We perform feature arithmetic by subtracting the text features of "bandage" and adding the text features of "snowman", and see a significantly higher similarity score of 0.62. Further improvement can be achieved by adding the text features of "hat" since the snowman wears a hat. See Table 8 for more qualitative examples.

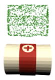
Feature Composition. We explore if our text features can capture compositional changes in the 3D space. We modify 3D models by altering object categories and counts. For each 3D modification, we also generate a corresponding modified text, as shown in the first three columns of Table 9. By testing similarities between the modified 3D model and modified text, we can measure how well our learned representations can generalize compositionally. Table 9 shows a quantitative comparison using 3D models of piano and stool. Across all tested compositional changes, MaskCL3D is the most effective among all methods. To further test our model on these two settings, we build a new dataset with 500 examples. The captions involve arithmetic and composition of diverse object shapes, colors, categories, sizes, and quantities. The results of 3D-to-text similarity of different

[1] 3D-text similarity metric is equivalent to the *negated* alignment metric up to a scale. We report alignment as squared distance here to be consistent with prior works on representation geometry (Wang & Isola, 2020).

Table 7: **Alignment and Uniformity of representations (Wang & Isola, 2020)**. MaskCL3D demonstrates superior performance in both Alignment and Uniformity metrics, indicating the effectiveness of the learned representations.

Method	**Alignment (↓)**	**Uniformity (↓)**
Point-CLIP (Zhang et al., 2022)$_{\text{[CVPR 2022]}}$	1.63	-0.52
ULIP (Xue et al., 2023a)$_{\text{[CVPR 2023]}}$	1.12	-1.25
ReCon (Qi et al., 2023)$_{\text{[ICML 2023]}}$	1.03	-1.53
OpenShape (Liu et al., 2023)$_{\text{[NeurIPS 2023]}}$	0.89	-1.76
MaskCL3D (ours)	**0.65**	**-2.79**

Table 8: **Quantitative comparisons of text features arithmetic.** The proposed MaskCL3D achieves more reasonable 3D-Text similarity scores across various arithmetic operations compared to state-of-the-art approaches.

3D Objects	**Text Arithmetic**	**ULIP** (Xue et al., 2023a)	**ReCon** (Qi et al., 2023)	**OpenShape** (Liu et al., 2023)	**MaskCL3D (ours)**
	a yellow bandage	0.45	0.43	0.39	**0.32**
	a yellow bandage − yellow + red	0.49	0.51	0.52	**0.57**
	a yellow bandage − yellow + white	0.52	0.53	0.56	**0.63**
	a yellow bandage − yellow + red + white	0.53	0.56	0.69	**0.82**
	a white bandage	0.23	0.21	0.18	**0.15**
	a white bandage − bandage + hat	0.39	0.36	0.31	**0.29**
	a white bandage − bandage + snowman	0.62	0.68	0.73	**0.82**
	a white bandage − bandage + snowman + hat	0.63	0.72	0.76	**0.87**

methods are reported in Table 10. Our method outperforms baselines on both arithmetic and composition tasks.

Text-to-3D Generation. The proposed representation can be used for 3D generation. We use the learned 3D representations for text-to-3D generation using Point·E and Shap·E. We evaluate the generated content using FID, CLIP Score, and CLIP R-Precision, and report the results of our method and Cap3D (Luo et al., 2023) in Table 11. Our method performs better on all metrics.

D. Related Work

Language-3D Pre-training. Language-3D pre-training aims to learn transferable representations of 3D objects given natural language supervision. Due to the limited amount of available text-3d pairs, previous methods (Zhang et al., 2022; Xue et al., 2023a) leveraged diverse pipelines based on language-2D pre-training to learn the alignment between 2D images and 3D objects. Typically, Point-CLIP (Zhang et al., 2022) generated multi-view 2D depth maps of 3D point clouds and aligned them with 3D class texts in zero-shot 3D object classification. Following up, object triplets from the three modalities (image, text, and 3D point clouds) were used in ULIP (Xue et al., 2023a) to learn the unified representations, where they rendered 60 multi-view 2D images from each CAD model for pre-training. More recently, OpenShape (Liu et al., 2023) adopted a multi-modal contrastive learning framework for representation alignment to capture multi-modal large-scale joint representations of large-scale text, image, and point clouds. Different from them, we do not require 4096 points for training and inference. Instead, we can support large-scale text-mesh pre-training on 3D objects using only 1024 points for efficient inference and also leverage masked point clouds for efficient training.

Contrastive Representation Learning. Contrastive representation learning has been explored in previous methods (Tian et al., 2020; Chen et al., 2020a;b; Grill et al., 2020; He et al., 2020; Chen et al., 2020c; Caron et al., 2020; Chen & He, 2021; Li et al., 2021; Wang et al., 2021; Mo et al., 2021; 2022) to be effective in learning discriminative representations. For example, SimCLR (Chen et al., 2020a) proposed the normalized temperature-scaled cross-entropy loss to pull away the features of each instance from those of all other instances in the training set by maximizing the similarity between positive samples and minimize the similarity between negative samples. For learning cross-modal representations, the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) model showcased the power of transferable image-language representations through zero-shot image classification. In this work, we leverage contrastive loss to align masked 3D representations with 2D image embeddings and text embeddings.

Masked Representation Learning. Masked representation learning aims to learn representations by reconstructing desired features of masked data given unmasked parts as clues, which has achieved promising results in natural language processing (Devlin et al., 2018; Liu et al., 2019; Sun et al., 2019; Conneau & Lample, 2019; Wettig et al., 2023) and computer vision (Bao et al., 2022; He et al., 2021; Wei et al., 2022; Xie et al., 2022; Chen et al., 2022; Wu & Mo, 2022; Feichtenhofer et al., 2022; Dong et al., 2023) community because of its faster speed and higher accuracy in downstream tasks, such as language-image masking in FLIP (Li et al., 2023b). Typically, BERT (Devlin et al., 2018) randomly masked 15% of word tokens and recovered them with unmasked words to learn generalizable textual features from a transformer (Vaswani et al., 2017a). To simplify the masked image encoding framework, MAE (He et al., 2021) reconstructed missing pixels of 75% masked patches using vision

Table 9: **Quantitative comparisons of text features composition.** The proposed MaskCL3D achieves more reasonable 3D-Text similarity scores across various composition types compared to state-of-the-art approaches.

3D Objects	Composition Type	Text Composition	ULIP (Xue et al., 2023a)	ReCon (Qi et al., 2023)	OpenShape (Liu et al., 2023)	MaskCL3D (ours)
 	Object categories	a stool	0.51	0.48	0.45	**0.39**
		a piano	0.63	0.65	0.69	**0.78**
	Object counts	two stools	0.43	0.39	0.33	**0.28**
		two pianos	0.61	0.59	0.56	**0.52**
	Object categories and counts	a piano and two stools	0.53	0.55	0.58	**0.63**
		a piano and a stool	0.69	0.73	0.79	**0.89**

Table 10: **Comparisons of features arithmetic and composition.**

Cases	ULIP (Xue et al., 2023a)	ReCon (Qi et al., 2023)	OpenShape (Liu et al., 2023)	MaskCL3D (ours)
Arithmetic	0.47	0.52	0.65	**0.83**
Composition	0.59	0.63	0.73	**0.87**

Table 11: **Comparisons of text-to-3D generation.** The proposed MaskCL3D achieves better FID, CLIP, and CLIP R-Precision scores compared to state-of-the-art approaches.

Method	Features	FID (↓)	CLIP Score	CLIP R-Precision (2k) R@1	R@5	R@10
Point·E	Cap3D	32.8	75.6	12.4	28.1	36.9
	MaskCL3D (ours)	**26.5**	**81.2**	**20.5**	**45.6**	**62.3**
Shap·E	Cap3D	35.5	79.1	20.0	38.8	47.3
	MaskCL3D (ours)	**28.6**	**85.2**	**29.7**	**57.3**	**73.5**

transformers (Dosovitskiy et al., 2021). More recently, researchers introduced diverse masking pipelines (Yu et al., 2022; Pang et al., 2022) to show the effectiveness of masked modeling in learning meaningful representations from point clouds. Different from them, we develop a simple yet effective framework to aggregate transferable representations of 3D objects with masked pre-training from natural language and 2D images together.

E. More Ablation Studies

Data Size. To investigate the impact of pre-training data size on the learned representations from our proposed MaskCL3D, we conducted additional experiments in this section. Specifically, we trained the same model (172 MB) on varying amounts of data, namely $46K$, $200K$, and $700K$. The results of the zero-shot recognition tasks on seven datasets are presented in Table 12. As the number of training data increases, the zero-shot performance increases as well. These improvements demonstrate the importance of using large-scale text-3D pairs from Caption-Objaverse in learning discriminative 3D and language representations.

Model Size. In Table 13, we also explore the effect of model size on the learned representations by training models with different numbers of parameters ($\{85$ MB$, 172$ MB$, 429$ MB$\}$) on $200K$ 3D-caption paired data. We can observe that when the number of parameters is 172 MB, we achieve the best performance in terms of all datasets. With the increase of model parameters from 85 MB to 172 MB, the proposed MaskCL3D consistently achieves better results on zero-shot recognition tasks. However, enlarging the parameters of the pre-training model from 172 MB to 429 MB decreases the performance on these benchmarks. This might be due to the limited number of pre-training pairs, which renders the large model challenging to capture more meaningful representations for zero-shot recognition.

Text to 3D Retrieval Visualization. To further evaluate the learned 3D and language representations, we provide text to 3D retrieval results of our method in Figure 4. Given a text description, our method is able to find the correct results from 46K objects. For example, given a caption of "a snowman with a hat", we successfully acquire top-5 ranking 3D objects with the same semantics as the descriptive texts. These examples further showcase the effectiveness of the proposed MaskCL3D in learning language-aware 3D representations from the large-scale text-3D pairs in Caption-Objaverse.

F. More Analyses of the Caption-Objaverse Dataset

In this section, we provide more analyses of the proposed dataset.

F.1. Statistics of Caption-Objaverse

We first calculate the number of words of each caption in the dataset and report the distribution of the number of captions to the caption length in Figure 5. We can observe that the distribution is close to a normal distribution between $2-19$. In particular, the length of captions primarily focuses on $7-12$. The total vocabulary size in Caption-Objaverse is $21,801$.

F.2. Dataset Examples

Figure 6 shows examples of caption-3D pairs randomly selected from Caption-Objaverse according to the length of captions, where we can generate high-quality descriptive

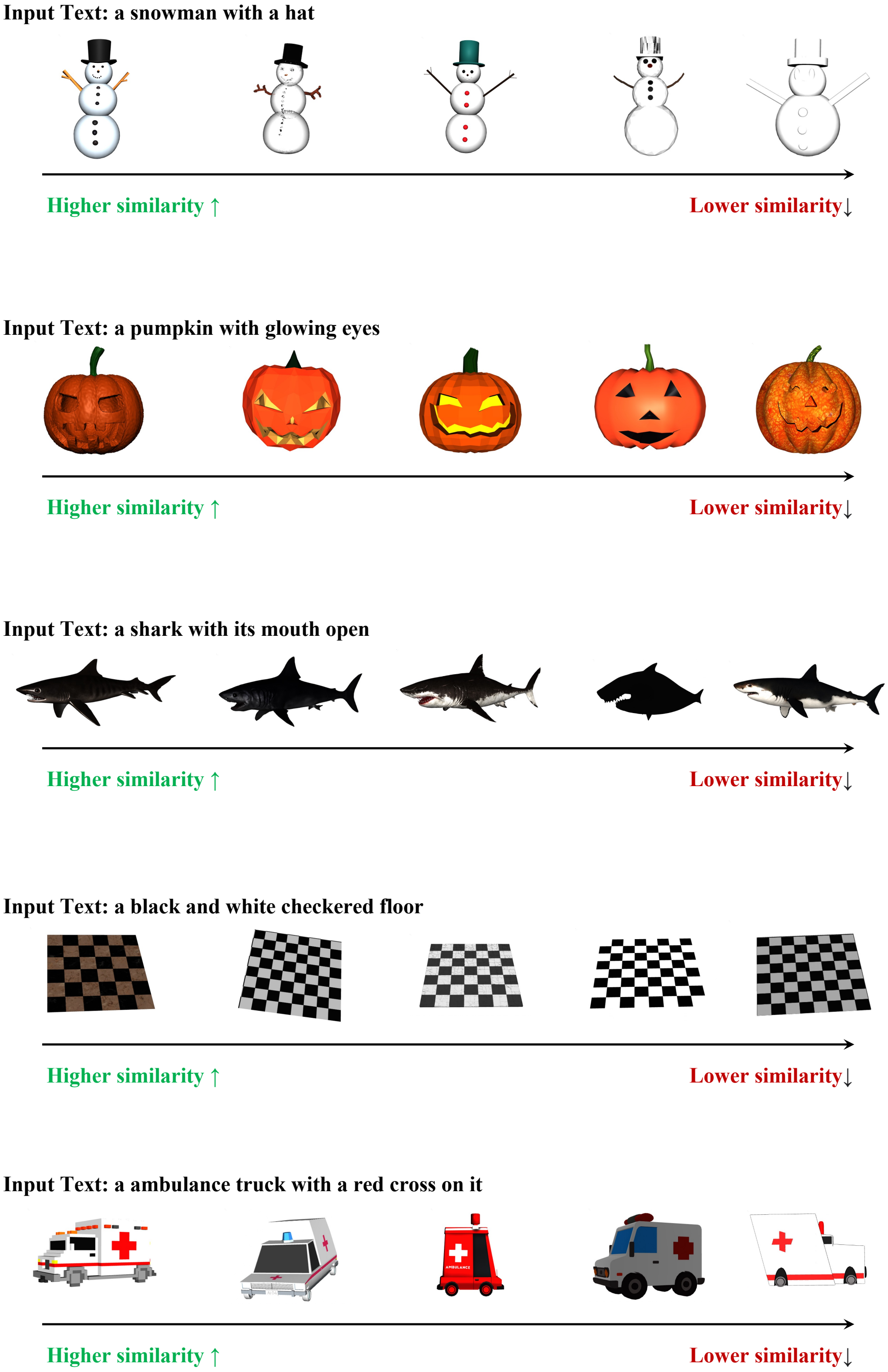

Input Text: a snowman with a hat
Higher similarity ↑
Lower similarity↓
Input Text: a pumpkin with glowing eyes
Higher similarity ↑
Lower similarity↓
Input Text: a shark with its mouth open
Higher similarity ↑
Lower similarity↓
Input Text: a black and white checkered floor
Higher similarity ↑
Lower similarity↓
Input Text: a ambulance truck with a red cross on it
Higher similarity ↑
Lower similarity↓

Figure 4: **Examples of text to 3D retrieval results in Caption-Objaverse.** Given a text description, our method is able to find the correct results from 46K 3D objects. The ranking results of top1 to top5 in text-to-3D retrieval are presented in a left-to-right sequence.

Table 12: **Ablation studies on training data size on zero-shot recognition tasks.**

Training Data Size	ModelNet 40	ModelNet 10	ScanObjectNN-ObjectOnly	ScanObjectNN-ObjectBg	ScanObjectNN-Hardest	ShapeNet 34	ShapeNet 55
46k	48.6	70.6	27.0	25.6	17.1	53.1	38.7
200k	50.1	71.6	28.2	26.5	17.9	54.5	40.3
700k	**50.6**	**72.1**	**28.7**	**26.9**	**18.3**	**55.1**	**40.6**

Table 13: **Ablation studies on model size (parameters) on zero-shot recognition tasks.**

Pre-train Model Params	ModelNet 40	ModelNet 10	ScanObjectNN-ObjectOnly	ScanObjectNN-ObjectBg	ScanObjectNN-Hardest	ShapeNet 34	ShapeNet 55
85 MB	48.3	69.8	26.2	23.6	16.3	51.2	38.2
172 MB	**50.1**	**71.6**	**28.2**	**26.5**	**17.9**	**54.5**	**40.3**
429 MB	49.6	70.9	27.6	25.8	17.3	53.9	40.1

texts on diverse 3D object models.

Figure 7 shows more examples of similar 3D objects in Caption-Objaverse. Our Caption-Objaverse dataset includes captions that provide specific details, enabling the distinction between similar 3D objects. For instance, given a 3D object of a chocolate donut, we generate the descriptive caption as “a chocolate donut with sprinkles on it”, which has more details than just a class annotation (*i.e.*, chocolate donut). Taking pizza as an example, we can generate captions “a slice of pizza with meat and vegetables on it” and “a pizza on a white surface with many small pieces of pepperoni” for two different types of pizzas. These examples further showcase the high quality of generated captions with paired semantics to each 3D object model.

G. Implementation Details

Our implementation is based on the PyTorch (Paszke et al., 2019) framework. For each 3D model, we sample 1,024 points and use a masking ratio of 50% for pre-training in the default setting. We also sample 20 views of 2D images from 3D mesh to extract 2D embeddings from a frozen ViT-L encoder in CLIP (Radford et al., 2021). We set the group size to c = 32 and the number of groups to k = 64 for FPS and KNN processing, following commonly-used settings in (Pang et al., 2022). For the 3D object encoder, the depth of self-attention layers is 12, the number of heads is 6, and the dimension of embeddings is 384. Following the prior work (Xue et al., 2023a), the depth of self-attention layers is 12, and the number of heads is 6 for the text encoder. Other text settings (context length and vocabulary size) follow CLIP (Radford et al., 2021). For the frozen language encoder, we use a pre-trained BERT (Devlin et al., 2018) base model to extract general textual features. The model is trained for 300 epochs using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of $1e-3$, a weight decay of 0.05, and a batch size of 128.

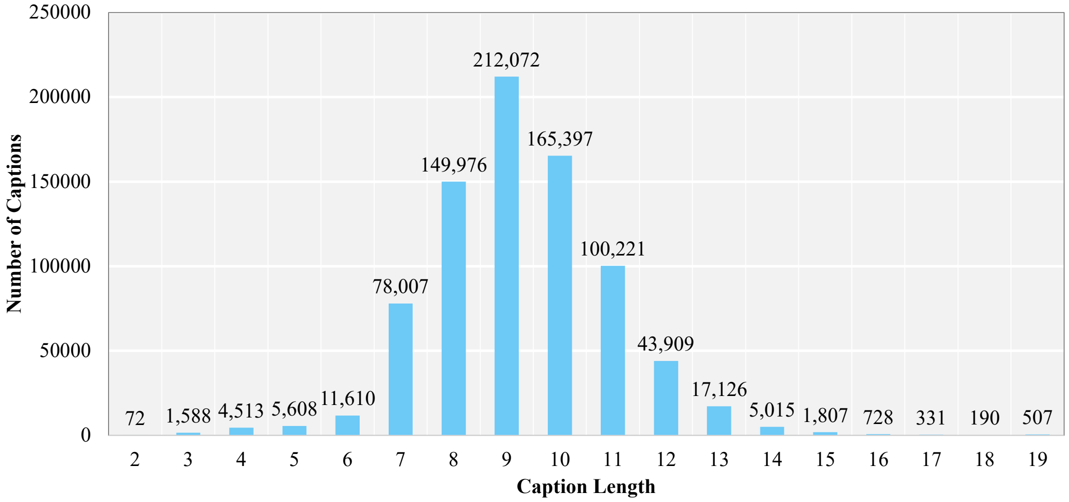


Figure 5: **Statistics of caption length in Caption-Objaverse.**

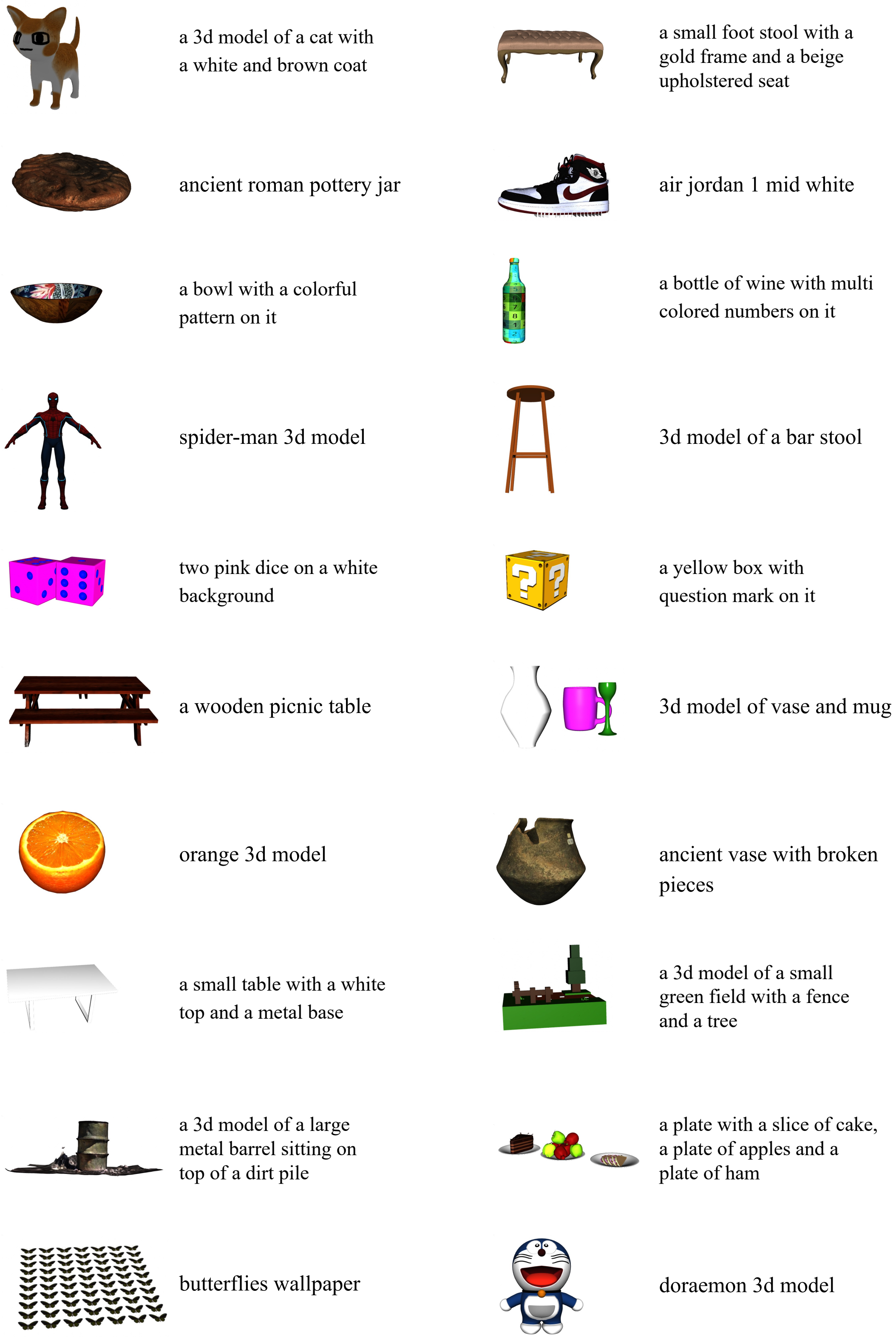

a 3d model of a cat with a white and brown coat
a small foot stool with a gold frame and a beige upholstered seat
ancient roman pottery jar
air jordan 1 mid white
a bowl with a colorful pattern on it
a bottle of wine with multi colored numbers on it
spider-man 3d model
3d model of a bar stool
two pink dice on a white background
a yellow box with question mark on it
a wooden picnic table
3d model of vase and mug
orange 3d model
ancient vase with broken pieces
a small table with a white top and a metal base
a 3d model of a small green field with a fence and a tree
a 3d model of a large metal barrel sitting on top of a dirt pile
a plate with a slice of cake, a plate of apples and a plate of ham
butterflies wallpaper
doraemon 3d model

Figure 6: **Examples of caption-3D pairs randomly selected from Caption-Objaverse.**

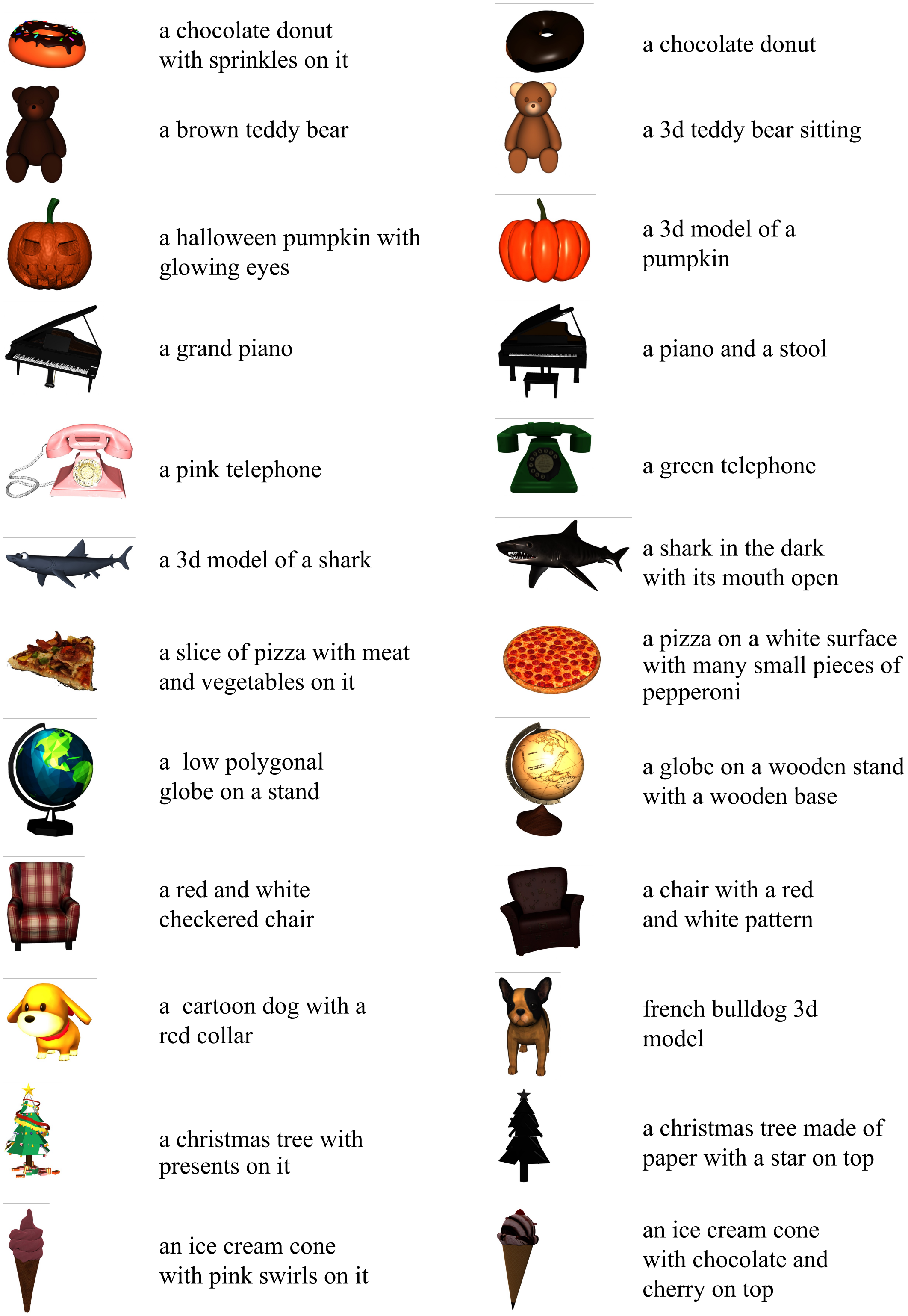

a chocolate donut with sprinkles on it
a chocolate donut
a brown teddy bear
a 3d teddy bear sitting
a halloween pumpkin with glowing eyes
a 3d model of a pumpkin
a grand piano
a piano and a stool
a pink telephone
a green telephone
a 3d model of a shark
a shark in the dark with its mouth open
a slice of pizza with meat and vegetables on it
a pizza on a white surface with many small pieces of pepperoni
a low polygonal globe on a stand
a globe on a wooden stand with a wooden base
a red and white checkered chair
a chair with a red and white pattern
a cartoon dog with a red collar
french bulldog 3d model
a christmas tree with presents on it
a christmas tree made of paper with a star on top
an ice cream cone with pink swirls on it
an ice cream cone with chocolate and cherry on top

Figure 7: **Examples of caption-3D pairs in our Caption-Objaverse dataset.** Our Caption-Objaverse dataset includes captions that provide specific details, enabling the distinction between similar 3D objects.