ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding

Le Xue[1] * **Ning Yu**[1] **Shu Zhang**[1] **Artemis Panagopoulou**[1,3] **Junnan Li**[1]
Roberto Martín-Martín[4] **Jiajun Wu**[2]
Caiming Xiong[1] **Ran Xu**[1] **Juan Carlos Niebles**[1,2] **Silvio Savarese**[1,2]
[1] Salesforce AI Research [2] Stanford University
[3] University of Pennsylvania [4] University of Texas at Austin

Abstract

*Recent advancements in multimodal pre-training have shown promising efficacy in 3D representation learning by aligning multimodal features across 3D shapes, their 2D counterparts, and language descriptions. However, the methods used by existing frameworks to curate such multimodal data, in particular language descriptions for 3D shapes, are not scalable, and the collected language descriptions are not diverse. To address this, we introduce ULIP-2, a simple yet effective tri-modal pre-training framework that leverages large multimodal models to automatically generate holistic language descriptions for 3D shapes. It only needs 3D data as input, eliminating the need for any manual 3D annotations, and is therefore scalable to large datasets. ULIP-2 is also equipped with scaled-up backbones for better multimodal representation learning. We conduct experiments on two large-scale 3D datasets, Objaverse and ShapeNet, and augment them with tri-modal datasets of 3D point clouds, images, and language for training ULIP-2. Experiments show that ULIP-2 demonstrates substantial benefits in three downstream tasks: zero-shot 3D classification, standard 3D classification with fine-tuning, and 3D captioning (3D-to-language generation). It achieves a new SOTA of **50.6% (top-1)** on Objaverse-LVIS and **84.7% (top-1)** on ModelNet40 in zero-shot classification. In the ScanObjectNN benchmark for standard fine-tuning, ULIP-2 reaches an overall accuracy of **91.5%** with a compact model of only 1.4 million parameters. ULIP-2 sheds light on a new paradigm for scalable multimodal 3D representation learning without human annotations and shows significant improvements over existing baselines. The code and datasets are released at https://github.com/salesforce/ULIP.*

1. Introduction

3D visual understanding has seen a surge of interests in recent years [8, 10, 21, 23, 35, 51] due to its growing applications in augmented reality and virtual reality (AR and VR) [2, 24, 27, 39, 45], autonomous driving [20, 54], and robotics [3, 48]. Despite this, the collections and annotations of 3D data remain a costly and labor-intensive process [4, 42, 50]. In response to this challenge, researchers have turned to other more abundantly available modalities, *e.g.*,

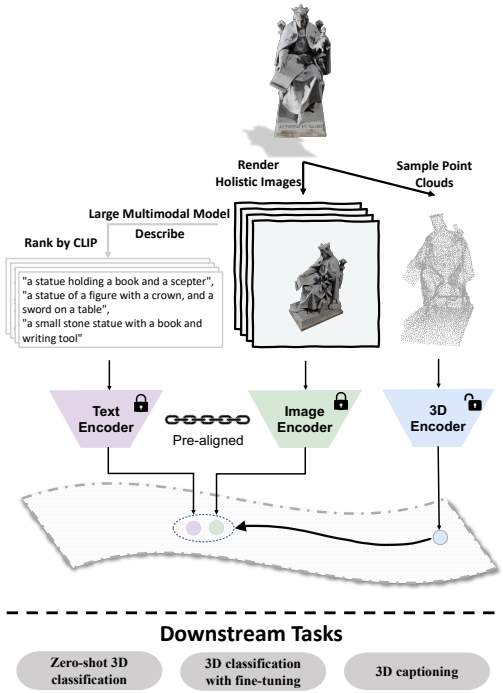

Figure 1. Overview of the ULIP-2 pre-training framework and its downstream tasks. The above part is the ULIP-2 pre-training framework, ULIP-2 employs a large multimodal model to automatically generate detailed descriptions for each 2D-rendered image from holistic viewpoints of a 3D shape. ULIP-2 takes advantage of a pre-aligned and frozen vision-language feature space to achieve alignment among the triplet modalities: holistic texts, images, and 3D point clouds. After the pre-training, the 3D encoder will be used in the downstream tasks. As shown in the figure, only the 3D data is required for this pre-training process.

* Contact: lxue@salesforce.com

image and natural language, to provide supervisory signals for learning 3D representations. This approach has not only led to improved unimodal representation but also cultivated a richer multimodal representation capability. The results have been promising, and to some extent, have alleviated the need for single-modal dense annotations in the 3D domain.

However, multimodal learning frameworks in this direction commonly face the challenge of assembling scalable, high-quality, and well-aligned multimodal data for 3D applications. We identify the language modality for 3D data as the critical bottleneck in this process. Existing frameworks tend to utilize manually annotated category names and short descriptions derived from metadata as the language counterparts for the 3D data. Those approaches [34, 52], however, lack scalability as they always rely on some degree of human annotations during the dataset collection process, which will be hard to scale up. Furthermore, existing methods are not comprehensive enough as the derived language information might not provide sufficient details and lacks variations, or appears to be noisy. This highlights the need for an innovative paradigm to provide language counterparts for 3D data that are both scalable and comprehensive, thus truly unleashing the potential of multimodal learning.

However, the optimal way to acquire and utilize language data for 3d modality is unclear. Although well-trained human annotators could potentially provide detailed language descriptions of 3D objects, such a method is both costly and lacks scalability. Moreover, identifying the appropriate language counterpart modality for a 3D shape is not a straightforward task.

To address these issues, we first reconsider what the 2D image counterpart modality for a 3D shape should be. Semantically, if we can render 2D images of a 3D shape from any viewpoint, the collection of all these rendered images should approximately encapsulate all information about this 3D shape, thus forming an appropriate image counterpart modality for 3D. By analogy, if we can linguistically describe a 3D shape from any viewpoint, the compilation of all these language descriptions from all perspectives should also approximately encompass all linguistically expressible information about this shape, thus forming an appropriate language modality for the 3D shape. In practice, for efficiency, we may sample a finite fixed set of holistic viewpoints instead of "any viewpoint". If we apply the same set of viewpoints for creating the language modality as we render the images, this task naturally boils down to describing the rendered 2D image for a given viewpoint. Leveraging the advances in large multimodal models that are trained on extensive language and image data, we utilize their ability to generate detailed language descriptions for the rendered images. This method allows us to automate the process in a scalable way as it only needs 3D data itself, while the rich knowledge from the large multimodal models is also distilled into the language descriptions. As a result, this automated and scalable strategy enriches the language modality with detailed, holistic descriptions, further aiding multimodal 3D representation learning.

In light of the preceding reasoning, and also in response to the challenge of scalable and comprehensive multimodal 3D data acquisition, we introduce ULIP-2, a novel framework that encompasses an innovative approach to generate well-aligned, holistic multimodal data for 3D understanding, coupled with an efficient multimodal pre-training architecture capable of robustly aligning these multimodal data, thereby harnessing the full potential of multimodal learning.

Given a 3D shape, our initial step involves extracting 3D point cloud data to serve as the 3D modality input. We then render this shape into a series of images from a fixed set of holistic viewpoints, providing the 2D modality input. For each rendered image, we employ a large multimodal model to generate a list of detailed descriptions, thereby establishing the language modality (as illustrated in Figure 2). This approach allows us to create scalable multimodal data for 3D, as it only requires the 3D data itself. Furthermore, by generating descriptions from a comprehensive set of holistic views, we address the prior issues of detail and comprehensiveness in the language modality. By employing an efficient multimodal pre-training architecture to align this multimodal data, we facilitate the learning of a comprehensive multimodal 3D representation, as described in Figure 1. Consequently, ULIP-2 offers a promising solution for scalable and comprehensive multimodal pre-training for 3D representation learning.

ULIP-2 advances beyond its predecessor, ULIP, by (1) proposing a manual-effort-free data creation paradigm for comprehensive multimodal learning, (2) leveraging this scalable paradigm to extend multimodal 3D learning to larger datasets, while scaling up both the vision-language and 3D backbones and (3) when pre-trained on the same datasets, ULIP-2 delivers impressive improvements over ULIP on all downstream tasks.

Our paper has the following main contributions:

1. ULIP-2 facilitates scalable multimodal pre-training without the need for human annotations, making it applicable to any 3D dataset, even unlabeled. It relies solely on the 3D data, enabling broader applicability and ease of use.
2. ULIP-2 achieves significant advancements in multimodal representation learning. On the challenging open-world Objaverse-LVIS benchmark, ULIP-2 attains a top-1 accuracy of **50.6%**, surpassing current SOTA (OpenShape [22]) by significant **3.8%**, despite ULIP-2 has a simpler and more streamlined framework; for zero-shot classification on ModelNet40, ULIP-2 reaches **84.7%**, even outperforming some fully supervised 3D classification methods [50]. Furthermore, it secures an overall accuracy of **91.5%** on the ScanObjectNN benchmark with

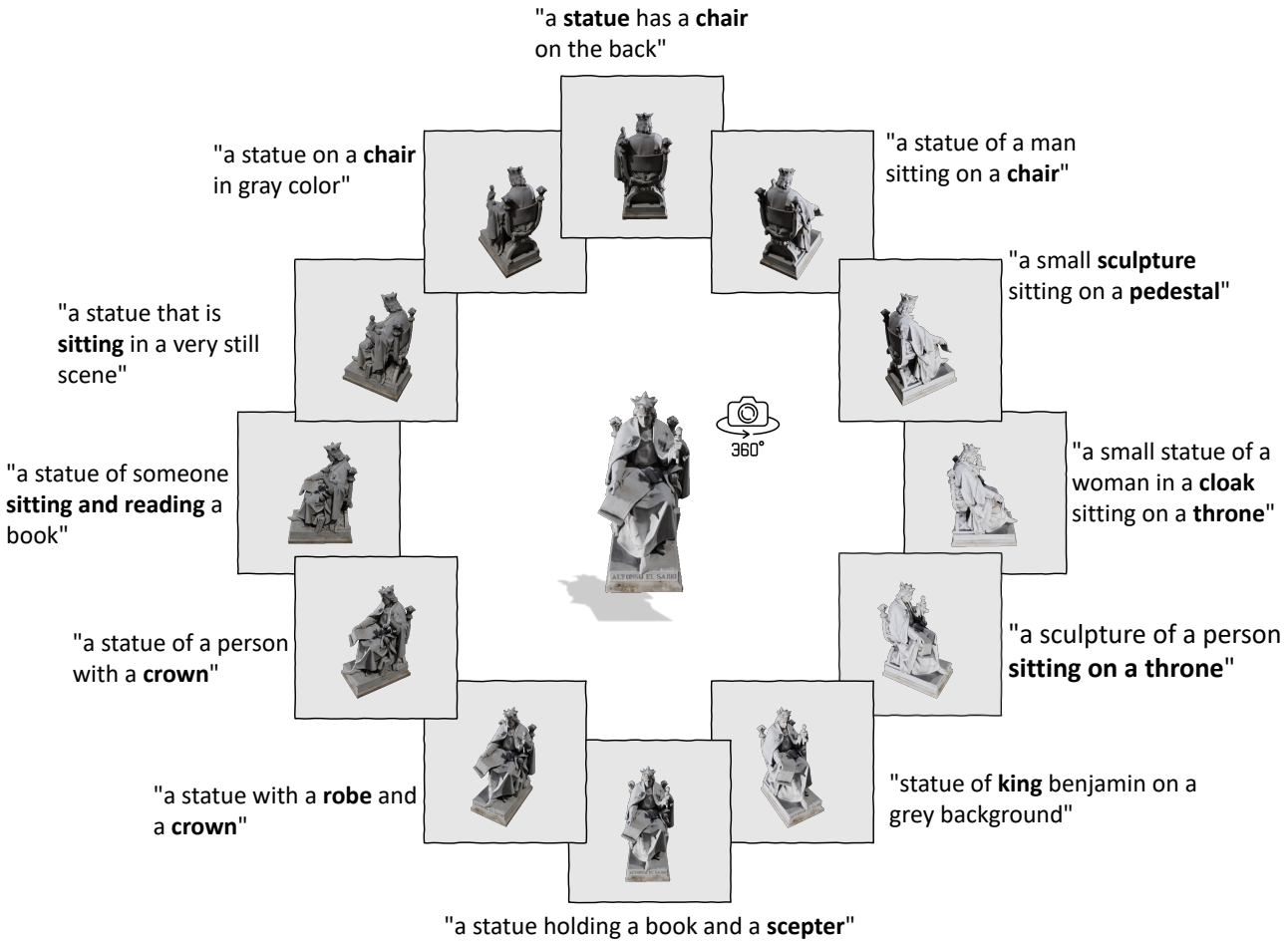

Figure 2. An illustration of language description generation from 2D images. These images are rendered from a set of holistic viewpoints of a 3D object. In some views, the chair is not visible, while in other views, the scepter/sword cannot be seen. Combining descriptions of all views is essential for the model to learn comprehensive and holistic information about the 3D object. From the metadata, the manual caption for this object is "Estatua de Alfonso X - José Alcoverro (1892)", which doesn't include much semantic information and could potentially harm the multimodal pre-training, unlike ULIP-2's holistic captions. More Samples can be found in the Appendix.

only 1.4 million parameters. The ULIP-2 encoder's 3D to language generation capabilities with LLMs are also demonstrated, highlighting its potential to keep pace with the growing LLM development. Moreover, ULIP-2 can effectively synergize with the ever-increasing capacity of 3D data and the development of large multimodal models.

3. We release two large-scale tri-modal datasets, **"ULIP-Objaverse"** and **"ULIP-ShapeNet"** triplets, consisting of point clouds, images, and language descriptions. The statistics of the datasets are detailed in Table 2.

2. Related Work

Multimodal Representation Learning. In recent years, multimodal representation learning has emerged as a popular research topic due to its remarkable capabilities and applications. Most research works focus on learning multimodal representation for only two modalities: language and image modalities, which have led to remarkable outcomes. One line of research in this area emphasizes the interaction between image regions and caption tokens using Transformer-based architectures [12, 15, 16, 30, 56], which exhibit strong predictive capabilities but are computationally expensive to train. Alternatively, methods such as CLIP [37] and SLIP [28] target generating single features for image and text independently and subsequently align these two modalities. This simplified architecture promotes robust and efficient large-scale pre-training, even on noisy data.

Recent works have demonstrated promising results by extending multimodal representation learning to 3D modality. ULIP [52] is one of the pioneering works in creating (3D point cloud - image - language) triplets, By aligning these three modalities together, ULIP enhances 3D represen-

tation learning and mitigates the need for single-modal dense 3D data annotations, thereby partially alleviating the data scarcity issue in 3D. A recent work [59] seeks to learn 3D representations from pre-trained 2D encoders via Image-to-Point Masked Autoencoders. However, this approach does not involve alignment with the language modality, which potentially limits its capacity for more complex multimodal tasks. The concurrent work of [22] further extends ULIP's framework to achieve stronger performance, but it still relies on manual annotation of 3D data and a complicated data engineering framework. However, ULIP-2 demonstrates that, with the proposed much simpler and streamlined framework, it can still achieve SOTA results on the challenging Objaverse-LVIS benchmark and outperform OpenShape by an impressive 3.8% in Objaverse-LVIS top-1 accuracy.

Despite the development of methods such as ULIP to reduce the single-modal dense annotation effort, they [22, 34, 52] still confront scalability challenges due to their dependency on dataset metadata and category names for obtaining the language counterpart modality. Additionally, the prompt-based pseudo-captions generated by these methods lack the fine-grained details, and variations that are necessary for comprehensive understanding. In contrast, ULIP-2 overcomes these limitations by leveraging the power and knowledge of state-of-the-art large multimodal models. This approach fundamentally diminishes data requirements and enriches the pre-train multimodal data, thereby enabling more efficient applications on larger datasets and yielding much stronger 3D representations.

Generative Large Multimodal Models. The expansion of transformer models, from GPT to GPT-4 [29, 43], demonstrates the effectiveness of scale in multimodal tasks. This approach, originating from [1], has seen considerable advancements in text generation from images [5, 16–19, 25, 46, 60, 61]. Our study leverages BLIP-2 [17] to generate diverse annotations for 3D shapes, facilitating learning richer multimodal 3D representations. We also conduct an ablation study in 5.2 on the used large multimodal models which indicates that, ULIP-2 benefits from the advancements in large multimodal models, synergizing with the rapid improvements in the field.

3D Point Cloud Understanding. PointNet [32] is a pioneering work that processes 3D point clouds directly [33]. Building on this, PointNeXt [36] emerges as a light-weight, high-performance variant. In the realm of self-supervised pre-training for point clouds, Point-BERT [55] moves a significant step forward with its transformer-based architecture, showcasing notable performance in zero-shot classification tasks. In ULIP-2, we leverage both Point-BERT and PointNeXt as our 3D encoders to harness their strong capabilities.

3. Method

ULIP-2 assimilates the pre-training framework of ULIP and introduces a scalable and comprehensive multimodal triplet creation paradigm that not only eliminates the need for human annotations but also significantly improves the learned multimodal 3D representations. By merging ULIP's efficient multimodal pre-training with this scalable triplet creation method, ULIP-2 paves the way for large-scale pre-training that essentially operates in a pseudo-self-supervised manner. We demonstrate that this method effectively mitigates the data scalability issue, and simultaneously advances the field of 3D representation learning to a new level of performance.

3.1. Preliminary: ULIP

ULIP [52] presents an efficient multimodal pre-training framework that constructs triplets encompassing three modalities: (1) the 3D modality, obtained by extracting 3D point cloud data; (2) the image modality, generated by rendering images from 3D shapes across multiple viewpoints; and (3) the language modality, derived by prompting dataset metadata such as descriptive terms and category names into cohesive sentences.

ULIP utilizes the ViT-B encoders from SLIP [28], a pre-trained vision-language model and a variant of the CLIP model, to learn 3D representations. It accomplishes this by aligning 3D modality features to the feature space shared by language and image modalities. ULIP-2 shares a similar objective with ULIP in aligning the (image, text, 3D) modalities, which prompts us to adopt its pre-training framework. Given the close resemblance in setup between ULIP and ULIP-2, we choose ULIP as our experimental baseline.

3.2. Scalable Triplet Creation

In ULIP-2, the model similarly utilizes three input modalities, though it only requires the 3D object data itself. As depicted in Fig. 1, given a 3D object, we extract 3D point clouds from the surface as the input to the 3D encoder and generate images from various viewing angles. We then leverage BLIP-2 [18], a cutting-edge large multimodal model, to generate descriptive texts for each rendered 2D image. For each image, we generate a set of sentences, rank them using CLIP similarities, and aggregate the top-1 description to form the language modality in the triplet.

This scalable triplet creation approach facilitates dataset scaling, eliminating the need for dataset metadata collection and necessitating only the 3D data itself. Our method is capable of aligning 3D representations with holistic image-text pairs in any unannotated 3D dataset, thereby providing a more scalable and comprehensive solution.

3.3. Tri-modal Pre-training

ULIP-2 aligns the triplet of 3D point clouds, 2D rendered images, and comprehensive descriptions to a unified fea-

Model	Pre-train dataset	Pre-train method	Manual captions?	Objaverse-LVIS top-1	Objaverse-LVIS top-5	ModelNet40 top-1	ModelNet40 top-5
PointCLIP [58]	–	–	–	1.9	5.8	19.3	34.8
PointCLIPv2 [62]	–	–	–	4.7	12.9	63.6	85.0
ReCon [34]	ShapeNet	ReCon [34]	✓	1.1	3.7	61.2	78.1
CLIP2Point [11]	ShapeNet	CLIP2Point [11]	✗	2.7	7.9	49.5	81.2
Point-BERT [55]	ShapeNet	OpenShape [22]	✓	10.8	25.0	70.3	91.3
Point-BERT [55]	Objaverse(no LVIS) + ShapeNet	OpenShape [22]	✓	38.8	68.8	83.9	97.6
Point-BERT [55]	Objaverse + ShapeNet	OpenShape [22]	✓	46.5	76.3	82.6	96.9
Point-BERT [55]	Objaverse + ShapeNet + (2 extra)	OpenShape [22]	✓	46.8	77.0	84.4	**98.0**
Point-BERT [55]	ShapeNet	ULIP [52]	✓	2.6	8.1	60.4	84.0
		ULIP-2	✗	16.4	34.3	75.2	95.0
	Objaverse(no LVIS) + ShapeNet	ULIP [52]	✓	21.4	41.9	68.6	86.4
		ULIP-2	✗	46.3	75.0	84.0	97.2
	Objaverse + ShapeNet	ULIP [52]	✓	34.9	61.0	69.6	85.9
		ULIP-2	✗	**50.6**	**79.1**	**84.7**	97.1

Table 1. Zero-shot 3D classification on Objaverse-LVIS and ModelNet40. The highlighted lines of OpenShape are from the current SOTA approach. Our method surpasses the current state-of-the-art (SOTA) OpenShape in zero-shot 3D classification, achieving a 3.8% higher top-1 accuracy on Objaverse-LVIS, and demonstrating comparable performance on ModelNet40, despite using fewer pre-training datasets. A tick in the "Manual captions?" column means the pre-trained model leverages 3D captions that, to some degree, rely on manual efforts, while a cross means the opposite.

ture space. We adopt the largest version of encoders from OpenCLIP (ViT-G/14) [13] for most of our experiments and freeze it during the pre-training. The feature space, already pre-aligned by OpenCLIP, serves as the target space where we aim to integrate the 3D modality.

During tri-modal pre-training, given a 3D shape $\mathbf{O}$, we extract its 3D point cloud $\mathbf{P}$, randomly sample its 2D rendered image $\mathbf{I} \sim \mathsf{render}(\mathbf{O})$, with its BLIP-2 generated language description $\mathbf{T} \sim \mathsf{blip2}(\mathbf{I})$, where render is the 3D-to-2D rendering operation and blip2 is to query BLIP-2 [18] for image description. We then extract the image feature $\mathbf{f}^{\mathrm{I}} = E_{\mathrm{I}}(\mathbf{I})$ and text feature $\mathbf{f}^{\mathrm{T}} = E_{\mathrm{T}}(\mathbf{T})$ based on the pre-aligned and frozen image encoder E_{I} and text encoder E_{T} in OpenCLIP [13]. We target to train a 3D point cloud encoder E_{P} such that its 3D feature $\mathbf{f}^{\mathrm{P}} = E_{\mathrm{P}}(\mathbf{P})$ is aligned with its image and text features. We formulate the 3D-to-image alignment using the contrastive loss similar in spirit to CLIP [37]:

$$\mathcal{L}_{\mathrm{P2I}} = -\frac{1}{2}\sum_i \log \frac{\exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_i^{\mathrm{I}}/\tau)}{\sum_j \exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_j^{\mathrm{I}}/\tau)} + \log \frac{\exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_i^{\mathrm{I}}/\tau)}{\sum_j \exp(\mathbf{f}_j^{\mathrm{P}}\mathbf{f}_i^{\mathrm{I}}/\tau)}, \tag{1}$$

where i, j are the sampling indices, and τ is a learnable temperature parameter. The first term indicates that the dot product of the 3D feature and the image feature of the same sample should stand out among other products where the *image features* are from different samples. Likewise, the second term indicates that the dot product of the 3D feature and the image feature of the same sample should stand out among other products where the *3D features* are from different samples.

Similarly, we formulate the 3D-to-text alignment loss as:

$$\mathcal{L}_{\mathrm{P2T}} = -\frac{1}{2}\sum_i \log \frac{\exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_i^{\mathrm{T}}/\tau)}{\sum_j \exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_j^{\mathrm{T}}/\tau)} + \log \frac{\exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_i^{\mathrm{T}}/\tau)}{\sum_j \exp(\mathbf{f}_j^{\mathrm{P}}\mathbf{f}_i^{\mathrm{T}}/\tau)}. \tag{2}$$

Our final training objective is to train the 3D encoder E_{P} that minimizes the sum of the two contrastive alignment losses above:

$$\min_{E_{\mathrm{P}}} \mathcal{L}_{\mathrm{P2I}} + \mathcal{L}_{\mathrm{P2T}}. \tag{3}$$

3.4. Scaling Up the 3D Multimodal Learning

Recognizing the benefits of more powerful image and text encoders for learning more generalized multimodal 3D representations, we extend our exploration beyond the smaller ViT-B model, previously utilized in ULIP. Our experiments focus on upgrading this vision-language backbone in the tri-modal alignment framework. Additionally, we investigate scaling up the model size, while keeping the other settings unchanged. The effectiveness of these modifications is evaluated through zero-shot classification tasks on both ModelNet40 and Objaverse-LVIS datasets. See Table 9.

Modality	ULIP-Objaverse	ULIP-ShapeNet
Point Clouds	$\sim$ 800k	$\sim$ 52.5k
Images	$\sim$ 10 million	$\sim$ 3 million
Language	$\sim$ 100 million	$\sim$ 30 million

Table 2. Statistics of ULIP-Objaverse and ULIP-ShapeNet triplets.

4. Experiments

4.1. ULIP-Objaverse Triplets and ULIP-ShapeNet Triplets Creation

We extract triplets of 3D point clouds, images, and language descriptions based on two large-scale datasets of 3D shapes. The first dataset is Objaverse [6], the recently released and largest-scale realistic 3D dataset. It has $\sim$ 800K real-world 3D shapes, each of which is associated with metadata containing a "name" field. For each 3D shape, we use Blender [14] to render 12 images, spaced equally by 360/12 degrees. For each rendered image, we employ BLIP-2-opt6.7B in BLIP-2 [18] to generate 10 detailed descriptions independently, which are then ranked using CLIP-VIT-Large [37] image-text similarity score. Based on an ablation study in Sec. 5.4, we use the top 1 description as the language modality input. Following ULIP and OpenShape, we use 10k, 8k, and 2k points from each 3D shape to accommodate different downstream tasks. Our generated well-paired triplets of comprehensive descriptions, 2D rendering images, and 3D point clouds are released as **ULIP-Objaverse** triplets.

The second dataset is ShapeNet [4], a renowned synthetic dataset. We employ its publicly available subset which has $\sim$ 52.5K 3D shapes with 55 annotated categories. For each shape, we follow ULIP to sample 30 equally spaced view angles, for each view angle, we render an RGB image and a depth map. The image description generation method is the same as that in Objaverse. We release these triplets as **ULIP-ShapeNet** triplets. More implementation details and ablation studies are included in the Appendix.

4.2. Downstream Tasks

We use the ModelNet40 [4], Objaverse-LVIS [6], and ScanObjectNN [41] datasets to benchmark ULIP-2. ModelNet40 is a synthetic CAD model dataset. It contains $\sim$ 9.8k training samples and $\sim$ 2.5k testing samples. Objaverse-LVIS is a subset of the Objaverse dataset with human-verified category labels. It has $\sim$ 46k samples spanning $\sim$ 1.2k categories, which is suitable for more challenging open-world zero-shot 3d shape classification. ScanObjectNN is a real-world 3D dataset with $\sim$ 2.9k samples under 15 categories. We follow the same dataset setup and preparation protocols used in ULIP and OpenShape, ensuring consistency in our comparisons.

We conduct experiments on three downstream tasks: (1) the zero-shot 3D classification task involving multimodal inputs and (2) the standard 3D classification task involving a single modality and (3) the 3D captioning task involving 3D-to-language generation with LLMs.

Evaluation Metrics. We adopt the same evaluation metrics used in ULIP: top-1 and top-5 accuracy for the zero-shot 3D classification task; overall accuracy and class average accuracy for the standard 3D classification task. For the new downstream task, 3D-to-language generation, we follow X-InstructBLIP [40] and employ CIDEr [44] score to quantitatively evaluate the quality of generated captions.

Backbones. We pre-train ULIP-2 on two representative backbones: Point-BERT [55] is a transformer-based backbone that performs strongly in ULIP's zero-shot classification experiments. PointNeXt [36] is a work that proposes a lightweight backbone based on PointNet++ [33] and delivers promising results on the ScanObjectNN benchmark.

4.3. Comparisons to Baselines

Zero-Shot 3D Classification. We follow the same procedure as in ULIP and OpenShape for zero-shot 3D classification, and compare with existing zero-shot approaches, including [11, 22, 34, 52, 58, 62]. We present the zero-shot 3D classification results on both Objaverse-LVIS and ModelNet40 in Table 1. First, we observe that, when **pre-trained on the same datasets**, benefit from ULIP-2 pre-training, Point-BERT obtains significantly better results than those pre-trained with ULIP. Specifically, when pre-train both ULIP and ULIP-2 on ShapeNet, ULIP-2 outperforms ModelNet40 top-1 accuracy over ULIP by **14.8%** and outperforms Objaverse-LVIS top-1 accuracy by **13.8%**. If pre-trained on Objaverse (excluding LVIS samples) and ShapeNet jointly, ULIP-2 outperforms ULIP by **15.4%** on ModelNet40 top-1 accuracy and outperforms ULIP by **24.9%** on Objaverse-LVIS top-1 accuracy. These gains underscore the efficacy of our approach, particularly the ranked holistic-view language descriptions and scaling strategies that amplify pre-training representation capabilities. The comprehensive language descriptions generated by the large multimodal model encapsulate its knowledge from a vast amount of language and image data, thus enriching the semantic richness of 3D shape descriptions and bolstering the alignment between language and 3D modalities.

Moreover, with ULIP-2's simple and streamlined framework, it achieves performance outperforming pre-existing baselines including concurrent OpenShape [22], which is also the current SOTA (46.8% in Objaverse-LVIS top-1).

Standard 3D Classification. We adhere to ULIP and community protocols for standard 3D classification. We present 3D classification results on ScanObjectNN hardest set in Table 3. When pre-trained on Objaverse and ShapeNet jointly for the same 3D encoder architecture with both ULIP and ULIP-2 frameworks, we observe that ULIP-2 (using the Point-BERT backbone) improves the baseline method (with-

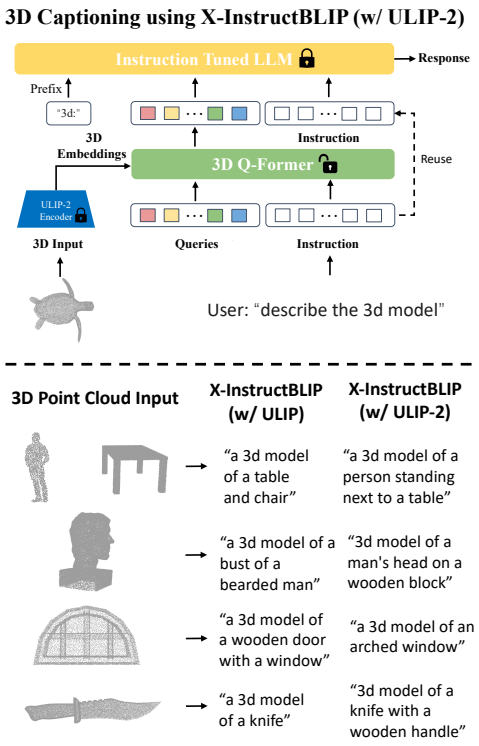

Figure 3. 3D-to-language multimodal generation using X-InstructBLIP framework [40].

Model	#Params (M)	Overall accuracy	Class-average accuracy
PointNet [32]	3.5	68.2	63.4
PointNet++ [33]	1.5	77.9	75.4
DGCNN [49]	1.8	78.1	73.6
MVTN [9]	11.2	82.8	–
RepSurf-U [38]	1.5	84.6	–
Point-MAE [31]	22.1	85.2	–
PointMLP [26]	12.6	85.7	84.4
Point-M2AE [57]	15.3	86.4	–
PointCMT [53]	12.6	86.7	84.8
ACT [7]	22.1	88.2	–
P2P [47]	-	89.3	–
Recon-s [34]	19.0	89.5	–
I2P-MAE [59]	12.9	90.1	–
Point-BERT (official)	22.1	83.1	–
Point-BERT (w/ ULIP)	22.1	88.7	–
Point-BERT (w/ ULIP-2)	22.1	89.7	–
PointNeXt (from scratch)	1.4	87.5	85.9
PointNeXt (w/ ULIP)	1.4	90.1	89.2
PointNeXt (w/ ULIP-2)	1.4	91.1	90.3
PointNeXt (w/ ULIP-2)*	1.4	**91.5**	**90.9**

Table 3. 3D classification results on ScanObjectNN. ULIP-2 significantly outperforms the baselines. * means the voting [26] is used.

out multimodal pre-training) by 6.6%. Using the PointNeXt backbone, ULIP-2 achieves a significant 4.0% performance gain over training from scratch, achieving an overall accuracy of 91.5% and establishing a new record on the ScanObjectNN benchmark with just 1.4 million parameters.

3D-to-Language Generation. As depicted in Figure 4, we adopt the X-InstructBLIP methodology [40] to integrate the ULIP-2 pre-trained encoder with a frozen large language model (LLM), endowing it with the capability to generate language from 3D data. For a fair comparison of 3D-to-language generation abilities, we plug frozen Point-BERT models into X-InstructBLIP's [40] framework, which are pre-trained under both ULIP and ULIP-2 frameworks with the same pre-training datasets (Objaverse + ShapeNet). Then we benchmark the 3D captioning abilities following [40]. All other variables are held constant during this evaluation. Captioning performance is measured using the PyCOCO-Tools Cider Score [44], offering a quantitative analysis of the models' captioning performance. Table 4 shows that ULIP-2 pre-trained encoder can significantly improve the captioning score by 28.3%, and Figure 4 shows qualitatively the generated captions using ULIP-2 pre-trained model is more accurate and descriptive.

Multimodal generation framework	Frozen 3D encoder	CIDEr score
X-InstructBLIP	PB w/ ULIP	132.2
X-InstructBLIP	PB w/ ULIP-2	**160.5**

Table 4. 3D-to-language generation using X-InstructBLIP [40], pre-trained on the same Objaverse + ShapeNet datasets setting. PB w/ ULIP-2 means Point-BERT pre-trained with ULIP-2 framework.

5. Ablation Study

5.1. Ablation on the effect of the generated captions

To ablate how the generated captions contribute to the performance, we conducted experiments aligned with the ULIP's settings, but with only one key modification: the language modality. Instead of using ULIP's manual descriptions, we utilized the top-1 ranked holistic-view captions generated by BLIP-2. Results in Table 5 show significant improvements in zero-shot classification on ModelNet40 when using these generated captions, demonstrating their crucial impact compared to the manual captions used in ULIP.

Pre-train language modality	ModelNet40 top-1	ModelNet40 top-5
Manual captions	60.4	84.0
Top-1 holistic BLIP-2 captions	**69.7**	**88.1**

Table 5. Point-BERT zero-shot 3D classification on ModelNet40, pre-trained on ShapeNet with SLIP ViT-B encoders (used in ULIP).

5.2. Different Large Multimodal Models

Considering that the language description quality from large multimodal models plays an important role in 3D represen-

tation pre-training, we conduct an ablation study over two such models. We use BLIP-2 [18] throughout the benchmarking experiments above. We hereby compare it to its earlier version BLIP [17] for the zero-shot 3D classification task using Point-BERT backbone pre-trained on ShapeNet. Results in Table 6 show that using BLIP-2 generated descriptions achieves better results than BLIP, thanks to its evolved image understanding capability, suggesting that as the large multimodal models advance, the performance of ULIP-2 can be expected to improve correspondingly.

Large multimodal models	top-1	top-5
BLIP [17]	67.7	88.6
BLIP-2 [18]	**69.7**	**88.8**

Table 6. Zero-shot 3D classification on ModelNet40. Pre-trained on ShapeNet with SLIP ViT-B encoders.

5.3. Number of 2D Views Per 3D Object

We further perform an ablation study for the zero-shot 3D classification performance w.r.t. the number of holistic views w/ its top-1 BLIP-2 caption in pre-training. Results in Table 7 demonstrate that, with the increase in the number of views, zero-shot classification accuracy increases accordingly. This validates our statement that diverse language descriptions of holistic views benefit multimodal 3D representation learning.

# Holistic views	Accuracy top-1	Accuracy top-5
1	54.8	77.9
2	58.1	80.5
15	69.3	88.6
30	**69.7**	**88.8**

Table 7. Zero-shot 3D classification on ModelNet40, pre-trained on ShapeNet with SLIP ViT-B encoders.

5.4. Top-k CLIP Ranked Captions Per 2D View

To assess the effectiveness of our top-1 BLIP-2 caption strategy, we conducted an ablation study on selecting different top-k of the 10 independently-generated BLIP-2 captions for the multimodal pre-training. The results in Table 8 indicate that, using the top-1 CLIP score ranked caption yields the best performance. This makes intuitive sense: the top-1 CLIP-scored caption tends to be more noise-proof, which is advantageous for multimodal learning in our context.

5.5. Scaling Up the Backbone Models

We examined the effectiveness of increasing the size of both the CLIP model and the 3D backbone model on performance. As Table 9 illustrates, a larger CLIP model improves results. For the 3D backbone, performance peaks at around 32.5M parameters, beyond which gains diminish. Therefore, this configuration is our chosen setup, balancing between performance and model size.

Top-k BLIP-2 captions selected	Accuracy top-1	Accuracy top-5
1	**69.7**	**88.8**
3	66.7	87.2
5	66.4	87.7
10	66.3	85.1

Table 8. ULIP-2 zero-shot 3D classification on ModelNet40, pre-trained on ShapeNet with SLIP ViT-B. For example, top-5 BLIP-2 captions selected means that in the pre-training, we will ensemble the top-5 CLIP CLIP ranked captions as the language modality.

CLIP size	3D encoder #Params(M)	ModelNet40 top-1	ModelNet40 top-5	Objaverse-LVIS top-1	Objaverse-LVIS top-5
ViT-B	21.9	71.4	89.7	28.3	52.6
ViT-G	21.9	76.3	94.1	35.0	62.5
ViT-G	5.3	75.0	94.7	34.1	61.1
ViT-G	21.9	76.3	94.1	35.0	62.5
ViT-G	32.5	77.0	94.0	35.7	62.9
ViT-G	43.1	76.8	94.8	35.9	62.6
ViT-G	85.7	76.5	94.7	35.9	62.7

Table 9. Zero-shot 3D classification on ModelNet40 and Objaverse-LVIS, all models are Point-BERT models which are pre-trained on Objaverse(no-LVIS). The highlighted gray line is the model setting we use to scale ULIP-2 to the larger Objaverse dataset. The smallest 5.3M model is used when only pre-trained on ShapeNet.

6. Conclusion and Discussion

We present ULIP-2, a novel framework for multimodal 3D representation learning. By leveraging large multimodal models for language description generation and scaling up the multimodal 3D pre-training, ULIP-2 not only addresses the quality and scalability challenges in existing multimodal 3D datasets but also demonstrates significant improvements in all downstream tasks. We also release "ULIP-Objaverse" triplets and "ULIP-ShapeNet" triplets, two large-scale tri-modal datasets to foster further research.

Limitations. ULIP-2's pre-training primarily utilizes object-level 3D shape datasets, which inherently differ from scene-level 3D data in their distribution and complexity. Exploring the application of the ULIP-2 framework to scene-level 3D data understanding, and leveraging the knowledge learned from object-level 3D data for this purpose, represents a compelling avenue for future research.

Broader Impact. ULIP-2 aims to minimize human annotation in 3D multimodal pre-training, reducing labor but potentially impacting low-skilled job markets. This dual impact, a common concern in AI advancements, underscores the need for broader considerations in AI research.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 4

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 1

[3] Cesar Cadena, Anthony R Dick, and Ian D Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. In *Robotics: Science and systems*, 2016. 1

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 6

[5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 4

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli Vanderbilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 6

[7] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022. 7

[8] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 1

[9] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2021. 7

[10] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 1

[11] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 6

[12] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 3

[13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 5

[14] Brian R Kent. *3D scientific visualization with Blender®*. Morgan & Claypool Publishers, 2015. 6

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3, 4

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4, 8

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR, 2023. 4, 5, 6, 8

[19] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *Annual Meeting of the Association for Computational Linguistics*, pages 2592–2607, 2021. 4

[20] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022. 1

[21] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7546–7555, 2021. 1

[22] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 2023. 2, 4, 5, 6

[23] Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, and Chunhong Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5239–5248, 2019. 1

[24] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 1

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 4

[26] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 7

[27] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 1

[28] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 3, 4

[29] OpenAI. Gpt-4 technical report. *OpenAI blog*, 2023. 4

[30] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023. 3

[31] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022. 7

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4, 7

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4, 6, 7

[34] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning (ICML)*, 2023. 2, 4, 5, 6, 7

[35] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. *arXiv preprint arXiv:2312.02980*, 2023. 1

[36] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *arXiv preprint arXiv:2206.04670*, 2022. 4, 6, 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 5, 6

[38] Haoxi Ran, Jun Liu, and Chengjie Wang. Surface representation for point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18942–18952, 2022. 7

[39] Manli Shu, Le Xue, Ning Yu, Roberto Martín-Martín, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. Model-agnostic hierarchical attention for 3d object detection. *arXiv preprint arXiv:2301.02650*, 2023. 1

[40] Anonymous Submission. X-InstructBLIP: A framework for aligning X-modal instruction-aware representations to LLMs and emergent cross-modal reasoning, 2023. 6, 7

[41] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 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)*, 2019. 6

[42] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 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*, pages 1588–1597, 2019. 1

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6, 7

[45] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 1

[46] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 4

[47] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *arXiv preprint arXiv:2208.02812*, 2022. 7

[48] Christian Wojek, Stefan Walk, Stefan Roth, and Bernt Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR 2011*, pages 1993–2000. IEEE, 2011. 1

[49] Bo Wu, Yang Liu, Bo Lang, and Lei Huang. Dgcnn: Disordered graph convolutional neural network based on the gaussian mixture model. *Neurocomputing*, 321:346–356, 2018. 7

[50] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2

[51] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 1

[52] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. *arXiv preprint arXiv:2212.05171*, 2022. 2, 3, 4, 5, 6, 1

[53] Xu Yan, Heshen Zhan, Chaoda Zheng, Jiantao Gao, Ruimao Zhang, Shuguang Cui, and Zhen Li. Let images give you more: Point cloud cross-modal training for shape analysis. *arXiv preprint arXiv:2210.04208*, 2022. 7

[54] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1

[55] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 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*, pages 19313–19322, 2022. 4, 5, 6, 1

[56] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 3

[57] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 7

[58] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Point-clip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 5, 6

[59] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *arXiv preprint arXiv:2212.06785*, 2022. 4, 7

[60] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6587, 2019. 4

[61] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, 2020. 4

[62] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-clip v2: Prompting clip and gpt for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 5, 6