PiSA: A Self-Augmented Data Engine and Training Strategy for 3D Understanding with Large Models

Supplementary Material

Overview

In the supplementary section, we first provide more related work and discussions to clarify existing solutions to self-augmented large language models. In addition, we provide more experimental details and results of PointLLM-PiSA. We organize our supplementary materials as follows:

- In Sec. A, we review more related work and provide more discussions.
- In Sec. B, we provide more details of data components.
- In Sec. C, we provide the results of 3D zero-shot classification task through the data processed by PiSA-Engine.
- In Sec. D, we clarify the prompts in PiSA-Engine and PiSA-Bench, as well as the human scoring criteria.

A. Related Work: 3D Object Understanding with Language

Through the integration of visual projections from point clouds, advanced models such as PointCLIP [15], PointCLIPv2 [17], and CLIP2Point [4] are enhancing 3D recognition capabilities by leveraging established CLIP [8] frameworks. ULIP [13] was the first to fuse 3D point cloud data with multimodal inputs from images and text, significantly advancing 3D representation learning and addressing some of the challenges posed by limited 3D-text data.

Further progress is exemplified by models like JM3D [10], CG3D [3], and Uni3D [16], which refine their algorithms by aligning point cloud encoders with CLIP embeddings using a triplet structure incorporating point clouds, images, and text. These approaches largely depend on internet-sourced data, such as Objaverse [2], which often includes short, noisy text annotations. Additionally, human annotations frequently fall short of conveying the full semantic richness present in 3D data.

Models like OpenShape [6] and ULIP-2 [14] utilize image-captioning algorithms to generate synthetic textual data, thus enhancing the quality of the triplet data. While OpenShape leverages multimodal language models, it restricts captioning to 2D images from fixed perspectives, potentially overlooking essential 3D attributes like depth and spatial relationships. Similarly, ULIP-2 employs BILP-2 [5] to generate detailed captions but limits the textual input to the top-k CLIP-ranked captions, which may omit critical semantic information necessary for precise 3D feature alignment during training. Consequently, these existing methods still fall short of achieving a comprehensive understanding of native 3D point clouds.

Number of all samples		*Proportion (%)*
Total	843,949	100
Brief-description Type	661,577	*78.39*
Detailed-description Type	15,055	*1.78*
Single-round Type	40,122	*4.75*
Multi-round Type	15,097	*1.79*
Instructions *(from PiSA-Engine)*	**112,098**	***13.28***

Table 1. **Statistics of 3D Instruction Following Data.**

B. More Details on Dataset Construction

Following the conventional approach for constructing instruction-following datasets, the original training data for PointLLM is generated by GPT-4. Specifically, GPT-4 is prompted to create a diverse range of instructions that align with captions described in the Cap3D [2] captions. This resulted in a large-scale dataset combining point-text instructions, comprising 660K brief descriptions and 70K detailed instructions.

For the training of PointLLM-PiSA, we extract 660K objects from the Cap3D [2] dataset. From this dataset, 3,000 objects are set aside exclusively for future testing, ensuring they are completely excluded from all stages of model training. Of these 3,000 reserved objects, 200 are used for testing in PointLLM, while the 40 objects in PiSA-Bench are also selected from this segregated set to prevent any information leakage.

Training data. Compared with 70K complex instructions in PointLLM, we further polish extra 112K data generated by PiSA-Engine for training. As shown in Table 1, despite our processed data accounting for only **13.28%** of the overall training dataset for PointLLM-PiSA, it delivers a significant improvement in performance.

C. 3D Zero-shot Classification

We first obtain method ULIP [13] pre-trained from ShapeNet [1] and the 3D-caption data processed by our PiSA-Enigne. As shown in Table 2, with templated-based prompt engineering, the accuracy of ULIP 59.56% is enhanced to 63.12% with a significant improvement of ***+3.56%***. Moreover, our PiSA-Enigne also enhances

Method	ModelNet40	
	Top1	Top5
PointCLIP	19.3	34.8
PointCLIPv2	63.6	85.0
ReCon[7]	61.2	78.1
CG3D	48.7	66.5
CLIP2Point	49.5	81.2
ULIP	59.56	83.95
+ ***PiSA-Engine***	**63.12** (↑ 3.56)	**85.45** (↑ 1.50)
ULIP-2	72.93	91.37
+ ***PiSA-Engine***	**73.58** (↑ 1.65)	**92.22** (↑ 0.85)

Table 2. **Zero-shot Classification on ModelNet40 [12].** We retest the ULIP family via open-source code for the sake of rigor.

Method	ScanObjectNN		
	Top1	Top3	Top5
OpenShape-SparseConv	56.7	78.9	88.6
OpenShape-PointBERT	52.2	79.7	88.7
ULIP-PointBERT	51.6	72.5	82.3
Uni3D-Ti	60.90	79.86	88.58
+ ***PiSA-Engine***	**61.49** (↑ 0.56)	**82.04**(↑ 2.18)	**89.72** (↑ 1.14)
Uni3D-B	63.88	82.73	90.27
+ ***PiSA-Engine***	**64.42** (↑ 0.54)	**83.63** (↑ 0.90)	**91.38** (↑ 1.11)
Uni3D-L	58.24	81.81	89.41
+ ***PiSA-Engine***	**59.15**(↑ 0.91)	**83.37**(↑ 1.56)	**90.87**(↑ 1.46)

Table 3. **Zero-shot Classification on ScanObjectNN [9].** *(Uni3D-Ti, Uni3D-B, and Uni3D-L represent Uni3D's tiny, base, and large versions, respectively.)*

ULIP-2 to achieve better performance, with an accuracy of 72.93% to 73.58%.

For Uni3D [16] pre-trained by the ensembled dataset, the *state-of-the-art* on zero-shot classification, we ensemble two predicted logits (the original one and the one processed through our method) by simple addition as the final output. Remarkably, with the integration of 3D-caption data processed by the PiSA-Engine, Uni3D-L of 58.24% is enhanced to 59.15% *(Top-1)* with an improvement of ***+0.91%*** for the challenging ScanObjectNN dataset, as shown in Table 3. This ensemble of two modalities highlights the complementary interactions between knowledge sources, enhancing overall model performance.

We further applied this ensembling method to additional datasets and observed similar improvements in performance. These experimental results confirm that the proposed method can serve effectively as a ***plug-and-play enhancement module*** for existing approaches, enabling robust point cloud understanding.

D. Additional Results

D.1. Data Filter Prompts

In the PiSA-Engine framework, 2D MLLMs such as Qwen2-VL [11] serve as supervisor, tasked with verifying and refining the responses generated by 3D MLLMs based on the provided images. This refinement process is detailed in Table 4, where the ***Input*** corresponds to the 3D caption produced by the 3D MLLMs, and the ***Output*** represents the enhanced response generated by Qwen2-VL.

D.2. PiSA-Bench Evaluation

Generative 3D Object Classification. In this task, GPT-4o acts as an evaluator to assess whether the model's response aligns with the object type described by the class in PiSA-Bench. The procedure is outlined in Table 5, where {ground_truth} represents the PiSA-Bench, and {model_output} refers to the response of the model. While the response of the model does not need to exactly replicate the ground truth class, it must correctly identify the object type.

3D Object Captioning. In this task, we utilize GPT-4o as an evaluator to compare captions generated by the model against PiSA-Bench, which serves as the ground truth. GPT-4o identifies elements from the PiSA-Bench captions and assesses the degree to which these elements are accurately or partially represented in the model-generated captions. Scoring is conducted on a scale of 0 to 100, with each identified element contributing equally to the overall score. The evaluation process is detailed in Table 6, where the {ground_truth} corresponds to the PiSA-Bench caption and the {model_output} refers to the caption generated by the model. Furthermore, the six aspects used for evaluation are explicitly defined and explained to GPT-4o to enhance understanding and accuracy, as outlined in Table 6. Table 7 and Table 8 present examples of positive and negative caption evaluations, respectively.

D.3. Human Scoring Criteria

Human evaluators are employed to assess captions in the object captioning task. To prevent bias, outputs corresponding to the same object from different models are grouped together and shuffled randomly. Evaluators then independently rate these captions while manually examining the colored visualization of the associated point cloud. Rather than employing captions generated by models as the ground truth, we rely on the native point cloud visualization to ensure more accurate and reliable evaluation results. The evaluation process involves assigning correctness score following these guidelines:

Correctness Score. Each distinct correct attribute in a model output is awarded one point, including category, color, shape, usage, material, relative position, spatial, and geometric information. For example, a white dinosaur correctly identified as a dinosaur and being white would receive two points. Partial correctness is graded on a scale of 0 to 1, depending on the degree of accuracy. For instance, if a model output described "a creature resembling a lizard" but

the object is specifically a dinosaur, it would be awarded 0.75 points.

Prompt	You are a helpful AI assistant. Now I will give you a description of a 3D model and several rendering images of this 3D model. You should correct the incorrect content that does not match the images, and refine this 3D description according to the given images in terms of its appearance and details. Do not edit depth, spatial, or relative position information: {model_output}
Sample 1	... ***Input:*** The 3D model represents a natural geological formation known as a cave. This cave is distinctly characterized by its many interconnected rooms and passageways, some of which are unoccupied. The cave's layout indicates several possible exploration paths, suggesting a level of intrigue and adventure often associated with spelunking activities. The cave model could be used in various digital scenarios such as gaming, film production, and geological studies, among others. ***Output:*** The 3D model represents a complex structure resembling a shipwreck. This shipwreck is characterized by its fragmented and damaged appearance, with various parts of the ship scattered and broken. The model includes elements such as wooden planks, metal structures, and debris, indicating a scene of destruction and abandonment. The shipwreck model could be used in various digital scenarios such as historical reconstructions, disaster simulations, or artistic projects.
Sample 2	... ***Input:*** This is a model of a split air conditioner unit's top part, primarily coloured in white with a dark grey rectangular panel at the top. It is designed to expel heat from the conditioned space, making it essential for cooling purposes. The dark grey panel possibly indicates the unit's control panel or display unit. The model reflects a common type of split air conditioner, often used in home and office spaces for temperature control. ***Output:*** This is a model of a window air conditioning unit, primarily colored in white. The unit features a rectangular shape with a front grille that has horizontal slats for air intake and exhaust. There is a small control panel on the right side of the front grille, which likely includes buttons for operation and a digital display. The model represents a common type of window-mounted air conditioner, widely used in residential and commercial settings for cooling individual rooms. The design suggests functionality and ease of installation in windows.
Sample 3	... ***Input:*** This 3D model depicts a vibrantly animated, multi-colored train in motion. The train's various parts - including the engine, carriages, and wheels - are painted with an array of colors, giving it a playful, kid-friendly aesthetic. The movement of the train suggests that it's either pulling or pushing a number of passengers, indicating its use in passenger transportation, commonly seen in railway systems worldwide. ***Output:*** This 3D model depicts a single, stationary train carriage painted in a combination of red and yellow. The train has multiple windows along its sides and is designed for passenger transportation, as indicated by the number "2" on both ends of the carriage. The model does not show any movement or additional carriages, suggesting it represents a single unit rather than a full train in motion.

Table 4. **The Prompt and Samples of Data Filter in PiSA-Engine.** The data filter is told to refine the visual details of 3D caption while not to amend the 3D information. The placeholder {model_output} refers to the 3D caption.

Prompt	You are a helpful AI assistant. Now, I will give you an answer from the model and an answer from the label. All you need to do is focus on these two answers and figure out whether they are referring to the same general class, focusing on the class of object, not attributes such as color, shape, count, spatial or usage. Respond with 'T' if they refer to the same class and 'F' if not. Also, provide a brief rationale (no more than 20 words) for your judgment. Remember, the answer from the model refers to one of the answers from the label; even if the answer from the model refers to the subclass of one of the answers from the label, you should respond as 'T'. Your response should follow the judgement standard of the prompt I give. Firstly, I will give you two examples of answer pairs as well as their responses: Example1: Input: 1. wooden board, table, pottery 2. This is a 3D model of wooden table. Output: T#Both refer to a table. Example2: Input: 1. historical vehicle, pioneer wagon, covered wagon, prairie schooner 2. The 3D object model depicts a quaint, old-fashioned cart. The cart is entirely brown, with two sturdy wooden wheels for mobility. The main body of the cart is in the shape of a large, semicircular curve, made of wood and affixed to the wheels. This curved body extends backward, forming a simple, straight tail. Despite its simplicity, it reflects a nostalgic charm and could be used in settings like historical reenactments, antiquated transportation exhibits, or in visual media for a touch of old-world atmosphere. Output: F#One refers to a wagon, the other to a truck. Now, analyze the following: ***Input:*** 1. {ground_truth} 2. {model_output} ***Output:***
Sample 1	... ***Input:*** 1. awning, shelter, canopy, shade 2. This is a 3D model of a structure featuring a vibrant blue and white striped shade covering a rectangular bench underneath. The shade, with its alternating blue and white stripes, provides ample protection from the sun, hinting at its use in creating a comfortable, outdoor sitting area. The bench's sturdy structure implies it can support significant weight, making it a practical choice for both indoor and outdoor setups, such as in a patio, backyard, or even a poolside area. ***Output:*** T#Both refer to a shade or shelter structure. ✓
Sample 2	... ***Input:*** 1. space exploration vehicle, Mars rover, robotic rover, planetary exploration equipment. 2. The 3D object model is a sleek, black robotic cleaning machine. It is designed with a compact and modern structure, indicating its ability to navigate through tight spaces. This machine likely utilizes advanced technology to autonomously clean floors, carpets, or other surfaces. Its primary function is likely to eliminate manual labor and time spent on cleaning, providing a deep clean with precision and ease. This can be used in both domestic and commercial environments, making it a convenient and efficient tool for maintaining cleanliness. ***Output:*** F#One refers to a space exploration rover, the other to a robotic cleaning machine. 55

Table 5. **The Prompt and Samples of GPT-4o in Open-vocabulary Classification.** GPT-4o needs to analyze two sentences to determine if they refer to the same general object or concept, focusing on the type of object, not attributes such as color, size, or shape. The placeholders {ground_truth} and {model_output} refer to the PiSA-Bench and the response of model, respectively.

Prompt You are a helpful AI assistant. Now, I will give you an answer from the model and an answer from the label. All you need to do is to evaluate these two answers from six aspects separately:
1. "description": Giving a comprehensive description of the whole 3D model.
2. "color": Demonstrating the color attribute of the whole or the individual objects.
3. "shape": Demonstrating the geometric attribute of the whole or the individual objects.
4. "count": Counting the number of the whole or the individual objects.
5. "spatial": Understanding the spatial relations between multiple objects in the 3D model.
6. "usage": Demonstrating the usage or the production purpose of the 3D model.

For any aspect above, you should identify the aspects mentioned in the answer from the model and calculate the percentage of these aspects correctly mentioned or partially matched in the answer from the label. Remember the score is to evaluate how much the two answers match. When evaluating and comparing each criterion, do not take other criteria into consideration. Score from 0 to 100. Consider similar concepts and synonyms.
Your response should include the scores of the six criteria (description, color, shape, count, spatial, and usage score) in the order above. Remember all scores range from 0 to 100. Firstly, I will give you several answer pairs and their corresponding scores. Your response format should follow the example of the prompt I give:

Provide your score (0-100) in the format of below:
' Scores for each aspects: **[description score, color score, shape score, count score, spatial score, usage score]** '
For clarity, consider this example:

Label:
"description": "This 3D model displays a wooden rectangular platform adorned with various items on top. The setup features a sizable dark clay vessel or cauldron positioned on a miniature stand at one end of the platform. At the other end, a table-like structure supports books, miniature vases, and what seems to be a witch's hat. The entire arrangement is decorated with vibrant speckles or splashes of paint in red, green, and blue, creating a magical or whimsical feel. The wood surface shows clear plank marks and a textured finish.",
"color": "The base displays a wooden brown shade, decorated with vibrant speckles in red, green, and blue. The large vessel is dark gray. The table holds brown books, two small vases (one blue and one green), and a dark gray pointed hat. The table also features the same speckled pattern as the wooden platform.",
"shape": "The principal structure is a rectangular wooden platform. It includes a circular pot at one end and a rectangular table at the other.",
"count": "The arrangement includes several items: one large pot, one table, multiple books, at least two small vases, and one pointed hat.",
"spatial": "The clay vessel is situated at one end of the wooden platform. The table, bearing various items, is located at the opposite end. All elements are positioned on the wooden platform.",
"usage": "This 3D model can be used in video games, animations, or virtual settings to craft a scene with a magical or fantasy theme, perhaps associated with witchcraft or alchemy. It can also be used for reading, writing, and eating."

Model: "This model portrays a vivid scene of a cartoon-styled table, overflowing with a variety of objects. The table seems to be in use, showcasing a naturalistic depiction of a cluttered table in a domestic or workspace. Objects on it reflect common items like books, pencils, cups, etc. suggesting its functionality as a piece of furniture where different activities such as reading, writing, or drinking can be performed."

Output: Scores for each aspects: **[35,0,0,30,35,75]**

Now score the following:
Label: {ground_truth}
Model: {model_output}
Output:

Table 6. **The Prompt of PiSA-Bench in Object Captioning.** GPT-4o evaluates the response of the model by identifying aspects mentioned in the PiSA-Bench caption and calculating the percentage of aspects that are correctly or partially matched in the caption generated by the model. The placeholders {ground_truth} and {model_output} refer to the PiSA-Bench caption and the response of the model, respectively.

Sample 1

...

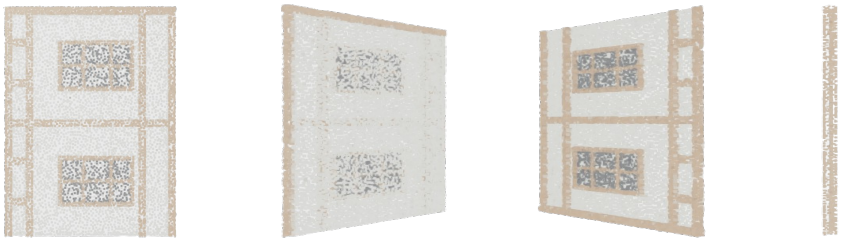
Ground Truth: "description": "This 3D model represents a whimsical unicorn pony, styled like those seen in 'My Little Pony'. It sports a light lavender body complemented by a purple mane and tail. The unicorn is adorned with a solitary horn and large, vivid eyes. The mane and tail display a dual-color design, predominantly purple with a stripe of mint green. On its side, it bears a small turquoise emblem or 'cutie mark'. The overall design of the model is smooth and rounded, typical of cartoon figures intended for a young audience.", "color": "The body is shaded light lavender. The mane and tail are mostly purple, accented with mint green stripes. The horn is pale pink. A small turquoise emblem is present on its side. The eyes are plain white, lacking iris or pupil in these images.", "shape": "The pony features a stylized, rounded body with slender legs, a petite snout, and an oversized head, typical of cartoon creatures. The mane and tail are designed with a flowing, wavy form. The horn exhibits a twisted shape.", "count": "There is one pony.", "spatial": "", "usage": "This 3D model is ideal for use in children's animated TV shows or movies, especially those akin to the 'My Little Pony' series. It could also be utilized in video games, mobile applications, or as a prototype for toy designs. Additionally, the model is suitable for digital artwork, 3D animation guides, or as a part of a character collection in a larger animated setting.

Model: A 3D model of a charming, small cartoon pony. The pony is primarily beige in color, with vibrant purple hair in its mane and tail. Its eyes are large and expressive, being white, which amplifies its cartoonish look. The size and design suggest it could be a character from a kid's show or an animated series, usually used for entertainment purposes like TV shows, toys, or animations.

Output: Scores for each aspect: **[60, 50, 75, 100, 0, 85]**
1. **Description**: 60 - The model captures the general idea of a cartoon pony with a colorful mane and tail, but misses details like the horn and cutie mark, and does not specify My Little Pony styling.
2. **Color**: 50 - The model identifies the purple hair correctly but describes the body color as beige instead of light lavender, and misses mint green stripes in the mane and the turquoise emblem.
3. **Shape**: 75 - The rounded body, large eyes, and cartoonish style are captured well, but there is no mention of the horn or the characteristic slender legs and snout described in the label.
4. **Count**: 100 - Both answers agree there is one pony.
5. **Spatial**: 0 - The model does not include any spatial details.
6. **Usage**: 85 - The model accurately describes the usage similar to the label but lacks specificity regarding My Little Pony series and other possible uses like digital artwork."

Table 7. **The Positive Sample of PiSA-Bench in Zero-shot Object Captioning.** *(The "spatial" aspect being null indicates that the 3D object is a single, cohesive entity, without any spatial relationships or information about separate components.)*

Sample 2



...

Ground Truth: "description": "This 3D object depicts a portion of a building wall, equipped with two windows. The color of the wall ranges from light blue to off-white, complemented by window frames and trim in shades of tan or light brown. Each window is divided into six panes configured in a 2x3 layout. Above these, a series of smaller, rectangular transom windows are placed. The model includes intricate details like the framing around the windows, paneling on the walls, and ornamental trim.", "color": "The primary wall color is either light blue or off-white, while the detailing around the windows and other architectural features are in tan or light brown. The glass of the window panes is a darker blue-gray.", "shape": "The model is shaped as a rectangle, featuring two main rectangular windows. It incorporates various geometric forms, such as the square and rectangular panes and the straight lines of the trim.", "count": "The model contains two primary windows, each with six panes, and six smaller transom windows positioned above the main ones.", "spatial": "", "usage": "This 3D model is suitable for use in architectural visualizations, interior design applications, video game settings that include building interiors or exteriors, or as a guide for building or remodeling projects."

Model: A rectangular flat board featuring a light hue, with a smooth texture. It is outfitted with a noticeable wooden structure, comprising of two vertical posts and two horizontal beams, creating a 'U' shape. This kind of object could be a component of a furniture item, like a cabinet or wardrobe, with the wooden structure providing support and organization benefits.

Output: Scores for each aspects: **[20, 25, 25, 20, 0, 30]**
1. **Description**: 20 - The model describes a rectangular board with a 'U' shape wooden structure, which aligns partially with the label's discussion of a wall section equipped with windows. However, the key architectural elements (windows, panels, ornamental trim) are missing.
2. **Color**: 25 - The model mentions a "light hue", partially matching the label's light blue/off-white description, but lacks specific details on other color features like tan/light brown window trim.
3. **Shape**: 25 - Both refer to a rectangular form, and there's mention of structures (the 'U' shape in the model); however, the detailed shapes of window panes and trims described in the label are not addressed.
4. **Count**: 20 - The model does not address the count explicitly but there is a mention of "two vertical posts and two horizontal beams", which could partially align with counting window frames but misses the specific panes.
5. **Spatial**: 0 - The model does not describe any spatial relations, whereas the label doesn't include spatial relationships either.
6. **Usage**: 30 - The model proposes usage related to furniture components, which is different from the architectural visualization purpose but relates to an object-based interpretation.

Table 8. **The Negative Sample of PiSA-Bench in Zero-shot Object Captioning.** *(The "spatial" aspect being null indicates that the 3D object is a single, cohesive entity, without any spatial relationships or information about separate components.)*

References

[1] 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. 2

[2] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[3] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M. Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition, 2023. 2

[4] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W. H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training, 2023. 2

[5] 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*, pages 19730–19742. PMLR, 2023. 2

[6] 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, 2023. 2

[7] 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*, pages 28223–28243. PMLR, 2023. 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[9] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[10] Haowei Wang, Jiji Tang, Jiayi Ji, Xiaoshuai Sun, Rongsheng Zhang, Yiwei Ma, Minda Zhao, Lincheng Li, Zeng Zhao, Tangjie Lv, and Rongrong Ji. Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023. 2

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[12] 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. 3

[13] 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

[14] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding, 2023. 2

[15] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip, 2021. 2

[16] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3

[17] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning, 2023. 2