Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following

Ziyu Guo[*1], Renrui Zhang[*†‡1,2], Xiangyang Zhu[2], Yiwen Tang[2], Xianzheng Ma[2], Jiaming Han[1,2]
Kexin Chen[1], Peng Gao[2], Xianzhi Li[‡3], Hongsheng Li[1], Pheng-Ann Heng[1]

[*] Equal contribution † Project leader ‡ Corresponding author

[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory
[3]Huazhong University of Science and Technology

{zyguo, pheng}@cse.cuhk.edu.hk, zhangrenrui@pjlab.org.cn

Abstract

*We introduce **Point-Bind**, a 3D multi-modality model aligning point clouds with 2D image, language, audio, and video. Guided by ImageBind, we construct a joint embedding space between 3D and multi-modalities, enabling many promising applications, e.g., any-to-3D generation, 3D embedding arithmetic, and 3D open-world understanding. On top of this, we further present **Point-LLM**, **the first** 3D large language model (LLM) following 3D multi-modal instructions. By parameter-efficient fine-tuning techniques, Point-LLM injects the semantics of Point-Bind into pre-trained LLMs, e.g., LLaMA, which **requires no 3D instruction data**, but exhibits superior 3D and multi-modal question-answering capacity. We hope our work may cast a light on the community for extending 3D point clouds to multi-modality applications. Code is available at https://github.com/ZiyuGuo99/Point-Bind_Point-LLM.*

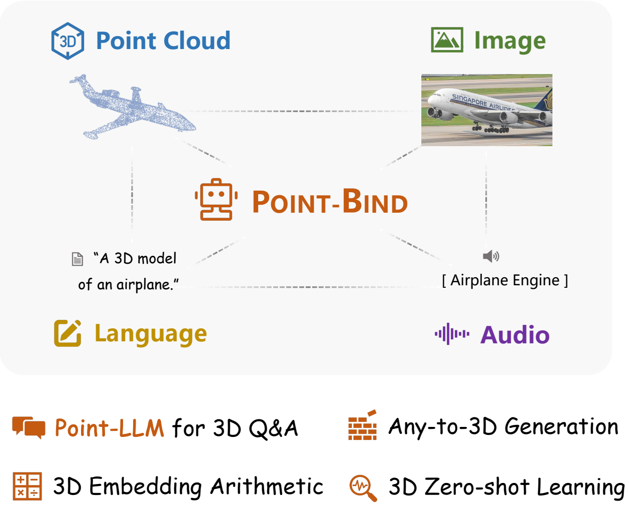


Figure 1. **Characteristics of Point-Bind.** We propose to align 3D with multi-modalities and develop a unified framework, Point-Bind, which extends various 3D multi-modal applications. Based on Point-Bind, we further introduce Point-LLM, a 3D large language model with bilingual 3D instruction-following capacity.

1. Introduction

In these years, 3D vision has gained significant attention and development, driven by the rising popularity of autonomous driving [11, 69, 71], navigation [72, 76, 98], 3D scene understanding [2, 43, 46, 74], and robotics [31, 67]. To extend its application scenarios, numerous efforts [1, 23, 92, 95] have been made to incorporate 3D point clouds with data from other modalities, allowing for improved 3D understanding [1, 23], text-to-3D generation [35, 49, 52], and 3D question answering [3, 28].

For 3D geometry understanding, previous works either leverage 2D-language embeddings to guide 3D open-world recognition [90, 100], or harness visual and textual semantics to assist 3D representation learning [39, 56, 84]. However, their perception capabilities are mostly constrained by limited modalities provided in the training phase. Inspired by 2D generative models [60, 63, 64], a collection of methods [35, 49, 52] has achieved text-to-3D synthesis with high quality and efficiency. Despite this, they lack the ability to generate 3D shapes conditioned on multi-modal input, i.e., any-to-3D generation. Another series of works connects descriptive natural language with 3D data, which is applied to 3D captioning [12, 87], question answering [3, 78], and visual grounding [24, 79]. Yet, they fail to utilize the pre-

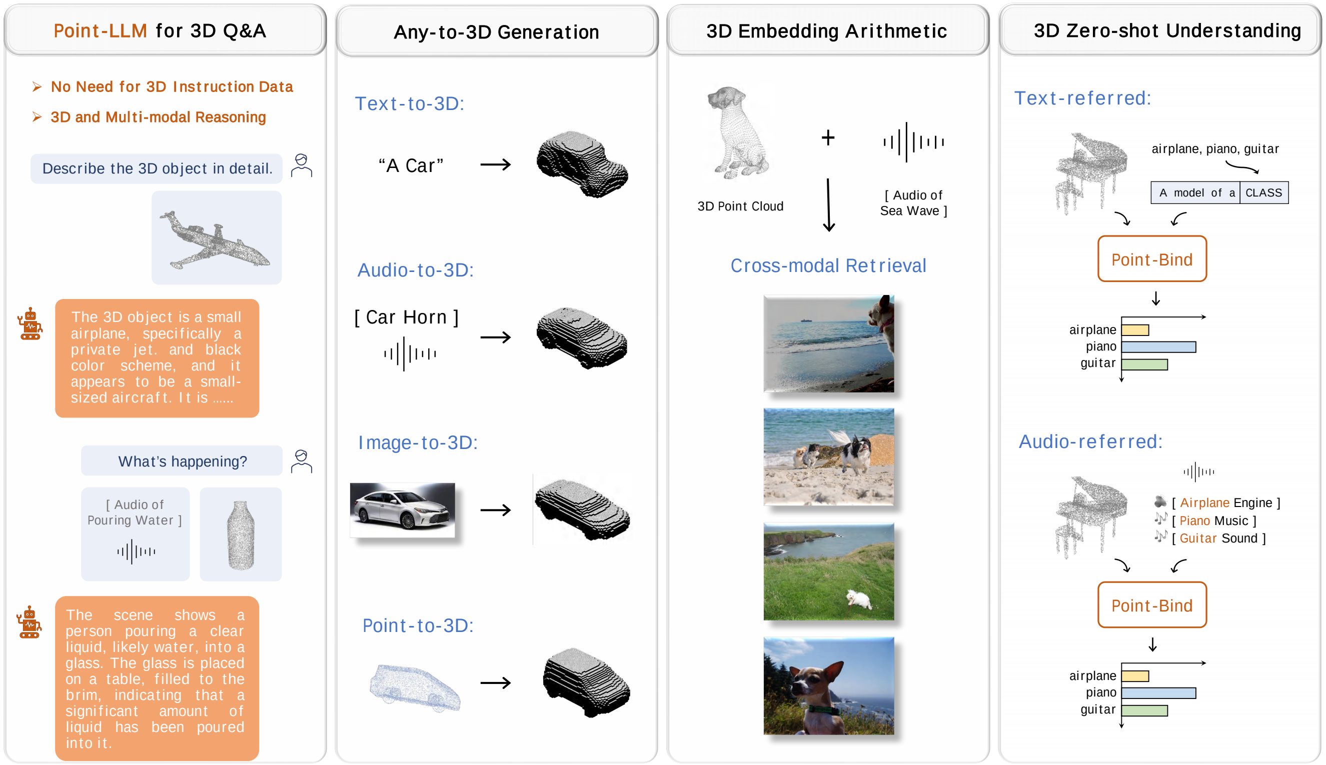

Point-LLM for 3D Q&A
No Need for 3D Instruction Data
3D and Multi-modal Reasoning
Describe the 3D object in detail.
The 3D object is a small airplane, specifically a private jet. and black color scheme, and it appears to be a small-sized aircraft. It is ......
What's happening?
[Audio of Pouring Water]
The scene shows a person pouring a clear liquid, likely water, into a glass. The glass is placed on a table, filled to the brim, indicating that a significant amount of liquid has been poured into it.
Any-to-3D Generation
Text-to-3D:
"A Car"
Audio-to-3D:
[Car Horn]
Image-to-3D:
Point-to-3D:
3D Embedding Arithmetic
3D Point Cloud
+
[Audio of Sea Wave]
Cross-modal Retrieval
3D Zero-shot Understanding
Text-referred:
airplane, piano, guitar
A model of a CLASS
Point-Bind
airplane
piano
guitar
Audio-referred:
[Airplane Engine]
[Piano Music]
[Guitar Sound]
Point-Bind
airplane
piano
guitar

Figure 2. **3D Multi-modal Applications of Point-Bind.** With a joint 3D multi-modal embedding space, Point-Bind enables many promising application scenarios, e.g., Point-LLM for 3D instruction following, 3D generation conditioned on any modalities, embedding-space arithmetic with 3D, and multi-modal 3D zero-shot understanding.

trained linguistic knowledge within large language models (LLMs) to better capture 3D geometrics.

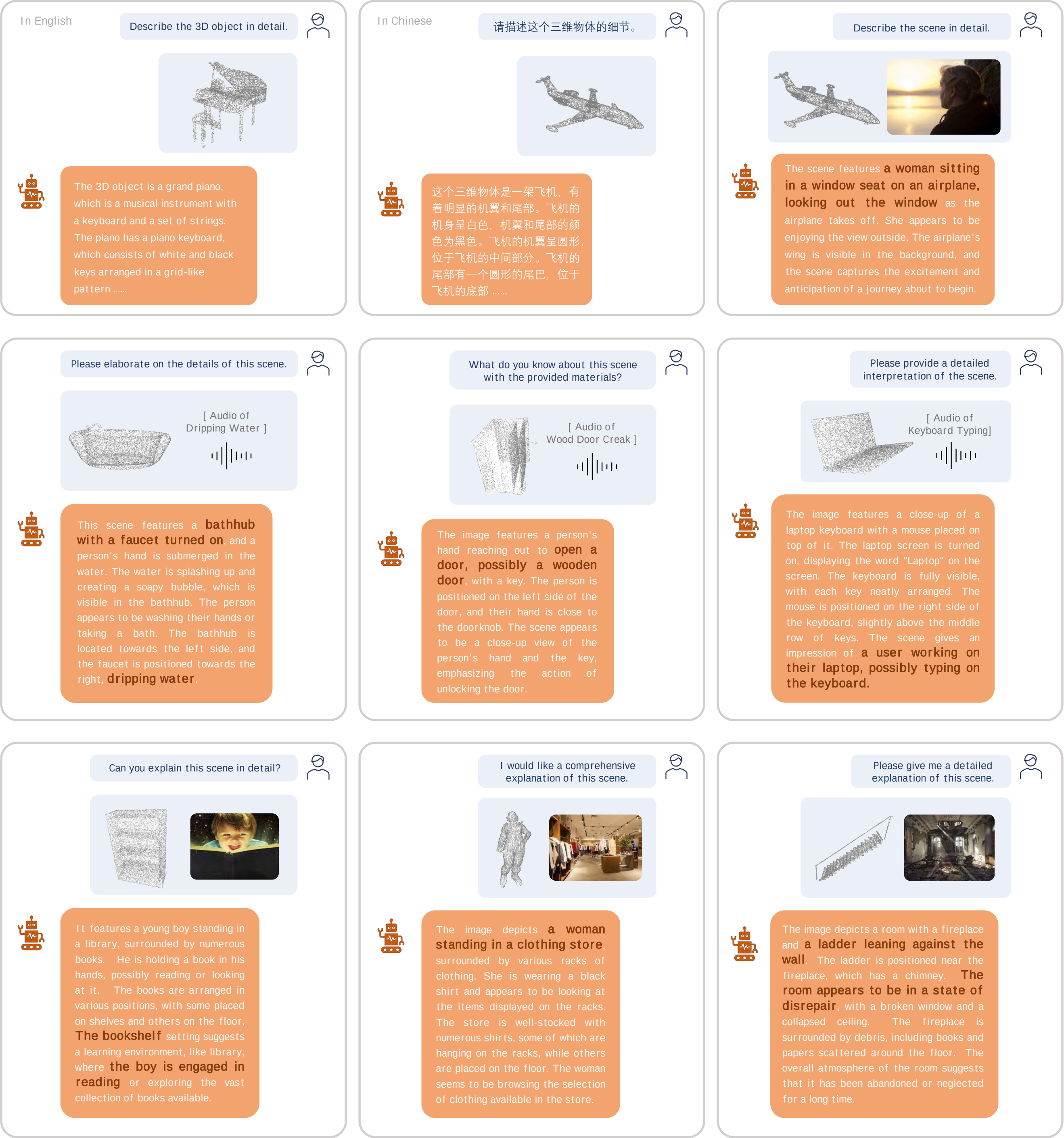
Therefore, how to develop a unified 3D framework aligning with multi-modality for general 3D learning still remains an open question. Very recently, ImageBind [22] was proposed to learn a shared representation space across six different modalities, i.e., images, text, audio, depth, thermal, and IMU data. Motivated by this, we ask the following question: ***can we construct a joint embedding space between 3D and multi-modality for unified 3D understanding, generation, and insturction following?***

To this end, we introduce **Point-Bind**, a 3D multi-modality framework that aligns point clouds with multiple modalities for general 3D analysis, as shown in Figure 1. Specifically, we collect 3D-image-text-audio pairs as the training data, and construct a joint embedding space guided by ImageBind. We adopt a contrastive loss between the extracted features from a trainable 3D encoder, e.g., I2P-MAE [92], and the frozen multi-modal encoders of ImageBind. Such a simple strategy can efficiently integrate different modalities into a unified representation space, and allows for various 3D-centric multi-modal tasks in Figure 2.

The main contributions of Point-Bind are as follows:

- ***Aligning 3D with ImageBind.*** Within a joint embedding space, Point-Bind firstly aligns 3D point clouds with multi-modalities guided by ImageBind, including 2D images, video, language, audio, etc.
- ***Any-to-3D Generation.*** Based on existing text-to-3D generative models, Point-Bind enables 3D shape synthesis conditioned on any modalities, i.e., text/image/audio/point-to-mesh generation.
- ***3D Embedding-space Arithmetic.*** We observe that 3D features from Point-Bind can be added with other modalities to incorporate their semantics, achieving composed cross-modal retrieval.
- ***3D Zero-shot Understanding.*** Point-Bind attains *state-of-the-art* performance for 3D zero-shot classification. Also, our approach supports audio-referred 3D open-world understanding, besides text reference.

Furthermore, on top of our joint embedding space, we propose to incorporate Point-Bind with LLaMA [73] to develop ***the first*** 3D large language models (LLMs), termed



Describe the 3D object in detail.

The 3D object is a grand piano, which is a musical instrument with a keyboard and a set of strings. The piano has a piano keyboard, which consists of white and black keys arranged in a grid-like pattern ……

请描述这个三维物体的细节。

这个三维物体是一架飞机，有着明显的机翼和尾部。飞机的机身呈白色，机翼和尾部的颜色为黑色。飞机的机翼呈圆形，位于飞机的中间部分。飞机的尾部有一个圆形的尾巴，位于飞机的底部 ……

Describe the scene in detail.

The scene features **a woman sitting in a window seat on an airplane, looking out the window** as the airplane takes off. She appears to be enjoying the view outside. The airplane's wing is visible in the background, and the scene captures the excitement and anticipation of a journey about to begin.

Please elaborate on the details of this scene.

This scene features **a bathhub with a faucet turned on**, and a person's hand is submerged in the water. The water is splashing up and creating a soapy bubble, which is visible in the bathhub. The person appears to be washing their hands or taking a bath. The bathhub is located towards the left side, and the faucet is positioned towards the right. **dripping water**

What do you know about this scene with the provided materials?

The image features a person's hand reaching out to **open a door, possibly a wooden door**, with a key. The person is positioned on the left side of the door, and their hand is close to the doorknob. The scene appears to be a close-up view of the person's hand and the key, emphasizing the action of unlocking the door.

Please provide a detailed interpretation of the scene.

The image features a close-up of a laptop keyboard with a mouse placed on top of it. The laptop screen is turned on, displaying the word "Laptop" on the screen. The keyboard is fully visible, with each key neatly arranged. The mouse is positioned on the right side of the keyboard, slightly above the middle row of keys. The scene gives an impression of **a user working on their laptop, possibly typing on the keyboard.**

Can you explain this scene in detail?

It features a young boy standing in a library, surrounded by numerous books. He is holding a book in his hands, possibly reading or looking at it. The books are arranged in various positions, with some placed on shelves and others on the floor. **The bookshelf** setting suggests a learning environment, like library, where **the boy is engaged in reading** or exploring the vast collection of books available.

I would like a comprehensive explanation of this scene.

The image depicts **a woman standing in a clothing store**, surrounded by various racks of clothing. She is wearing a black shirt and appears to be looking at the items displayed on the racks. The store is well-stocked with numerous shirts, some of which are hanging on the racks, while others are placed on the floor. The woman seems to be browsing the selection of clothing available in the store.

Please give me a detailed explanation of this scene.

The image depicts a room with a fireplace and **a ladder leaning against the wall**. The ladder is positioned near the fireplace, which has a chimney. **The room appears to be in a state of disrepair**, with a broken window and a collapsed ceiling. The fireplace is surrounded by debris, including books and papers scattered around the floor. The overall atmosphere of the room suggests that it has been abandoned or neglected for a long time.

Figure 3. **3D Question-answering Examples of Point-LLM.** Given 3D and multi-modal instructions, our Point-LLM can effectively generate detailed responses and conduct superior cross-modal reasoning. Notably, we do not need any 3D instruction data for training.

as **Point-LLM**. As shown in Figure 2, our Point-LLM can respond to language instructions with 3D point cloud conditions, and effectively capture spatial geometry characteristics. Referring to ImageBind-LLM [19], we utilize a bind network along with a visual cache model to bridge Point-Bind with LLaMA, and adopt zero-initialized gating mechanisms [20, 91] for parameter-efficient fine-tuning. With superior data efficiency, the entire training phase of Point-LLM ***requires no 3D instruction dataset***, and only utilizes public vision-language data [9, 13, 68, 70] for vision-language tuning. In this way, we enable LLMs to understand and conduct cross-modal reasoning for 3D and multi-modal data, achieving superior 3D question-answering capacity in both English and Chinese.

The main contributions of Point-LLM are as follows:

- ***Point-LLM for 3D Question Answering.*** Using Point-Bind, we introduce Point-LLM, the first 3D LLM that responds to instructions with 3D point cloud conditions, supporting both English and Chinese.
- ***Data- and Parameter-efficiency.*** We only utilize public vision-language data for tuning without any 3D instruction data, and adopt parameter-efficient fine-tuning techniques, saving extensive resources.
- ***3D and Multi-modal Reasoning.*** Via the joint embedding space, Point-LLM can generate descriptive responses by reasoning a combination of 3D and multi-modal input, e.g., a point cloud with an image/audio.

2. Related Work

Multi-modality Learning. Compared to single-modal approaches, multi-modal learning aims to learn from multiple modalities simultaneously, achieving more robust and diverse representation learning. Numerous studies have proved its efficacy, involving 2D images, videos, texts, and audio [15, 17, 48], and enhance the cross-modal performance for downstream tasks [5, 25, 37, 61], and video-text-audio integration for text generation [36]. The representative vision-language pre-training, CLIP [59], effectively bridges the gap between 2D images and texts, which encourages further exploration of cross-modality learning. Recently, ImageBind [22] successfully aligns six modalities in a joint embedding space, unleashing the power for emergent zero-shot cross-modal capabilities. However, ImageBind fails to investigate its efficacy on 3D point clouds. In the 3D domain, most existing cross-modal works introduce vision-language alignment [1, 10, 23, 84, 90] into 3D point clouds, and mainly focus on open-world recognition tasks, which ignore the potential of multi-modal semantics for wider 3D applications. In this paper, our Point-Bind develops a general 3D multi-modality model that aligns 3D point clouds with six other modalities guided by ImageBind, allowing for more diverse 3D cross-modal understanding.

Large Models in 3D. Large-scale pre-trained models have achieved remarkable downstream performance in language and 2D image processing. Inspired by this, many efforts have introduced 2D and language large models, to assist in 3D learning. The prior PointCLIP series [30, 90, 100] project 3D point clouds into depth maps, and utilize CLIP [59] for zero-shot recognition. Image2Point [82] instead converts 2D pre-trained models into 3D space as a good network initialization. By contrastive learning, ULIP series [84, 85] and other works [27, 39] pre-train 3D networks guided by the vision-language embedding space of CLIP. Another branch of work employs CLIP to guide the text-conditioned generation of 3D objects [32, 41, 65, 83] or stylized meshes [45, 47] by encoding descriptive textual input. Some works also adopt GPT-3 [6] to enhance the language-based understanding of 3D spatial geometry, such as PointCLIP V2 [100] and ViewRefer [24]. Different from them, we utilize ImageBind [22] to construct a joint embedding space between 3D point clouds and multiple modalities. The derived Point-Bind can well leverage the multi-modal semantics for general 3D cross-modal understanding, generation, and question answering.

Pre-training in 3D. In recent years, significant progress has been made in supervised learning for 3D vision tasks [54, 55, 57, 93, 99]. However, these approaches lack satisfactory generalization capabilities for out-of-domain data. To address this, self-supervised learning has emerged as a promising solution to enhance 3D transfer learning [10, 34, 53, 86]. Most self-supervised pre-training methods employ an encoder-decoder framework to encode point clouds into latent representations and then reconstruct the original data form [62, 66, 75]. Therein, Point-MAE [50] and Point-M2AE [89] introduce masked autoencoders [26] into 3D point clouds pre-training, achieving competitive results on different 3D tasks. Alternatively, cross-modal pre-training approaches are also leveraged to enhance the 3D generalization ability [40, 56, 58, 77]. For example, ACT [16] and I2P-MAE [92] utilize pre-trained 2D transformers as teachers to guide 3D representation learning. Inspired by previous works, we adopt collected 3D-image-text-audio pairs for self-supervised pre-training, and regard ImageBind's encoders as guidance for contrastive learning. In this way, the Point-Bind is pre-trained to obtain a joint embedding space between 3D and multi-modality, allowing for superior performance on different 3D downstream tasks.

3. Point-Bind

The overall pipeline of Point-Bind is shown in Figure 4. In Section 3.1, we first provide a preliminary of ImageBind [22]. Then, in Section 3.2 and 3.3, we elaborate on the training data and multi-modal alignment for Point-Bind, respectively. Finally, in Section 3.4, we introduce several 3D-centric applications derived from Point-Bind.

3.1. Preliminary of ImageBind

ImageBind [22] proposes an approach to combine multiple modalities together, which utilizes only image-paired

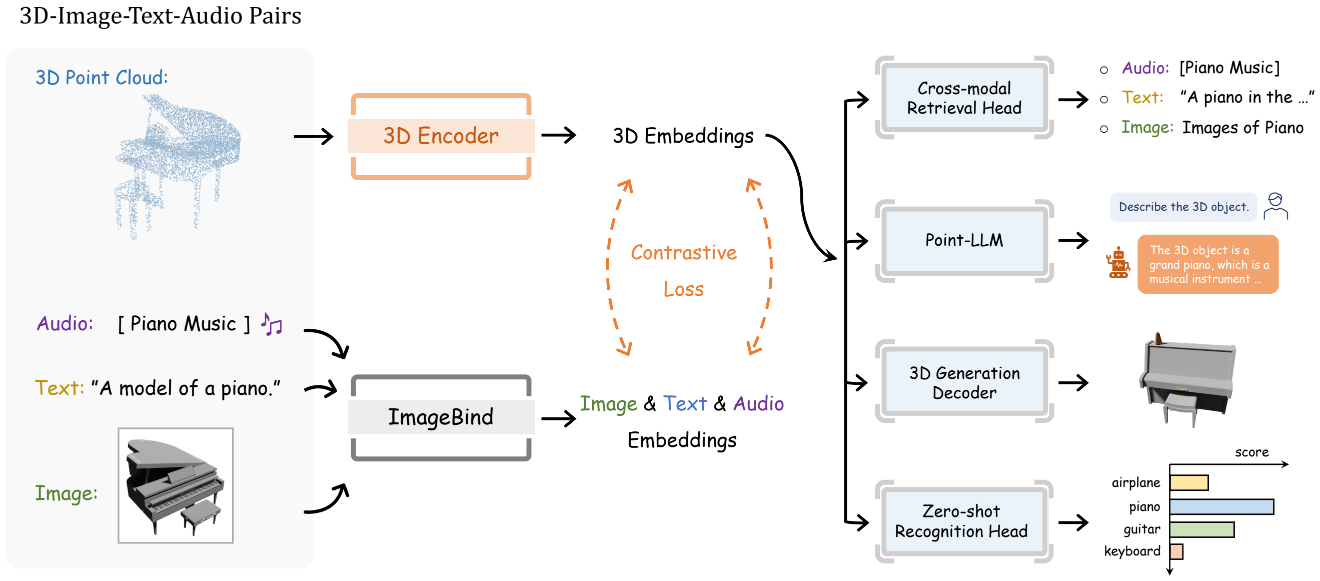

3D-Image-Text-Audio Pairs
3D Point Cloud:
Audio: [Piano Music]
Text: "A model of a piano."
Image:
3D Encoder
ImageBind
3D Embeddings
Contrastive Loss
Image & Text & Audio Embeddings
Cross-modal Retrieval Head
Audio: [Piano Music]
Text: "A piano in the ..."
Image: Images of Piano
Point-LLM
Describe the 3D object.
The 3D object is a grand piano, which is a musical instrument ...
3D Generation Decoder
Zero-shot Recognition Head
score
airplane
piano
guitar
keyboard

Figure 4. **Overall Pipeline of Point-Bind.** We collect 3D-image-audio-text data pairs for contrastive learning, which aligns 3D modality with others guided ImageBind [22]. With a joint embedding space, Point-Bind can be utilized for 3D cross-modal retrieval, any-to-3D generation, 3D zero-shot understanding, and developing a 3D large language model, Point-LLM.

data to learn a joint embedding space of six modalities, i.e., images, text, audio, depth, thermal, and IMU data. It does not need training dataset pairing all six modalities, but leverages the binding property of 2D images, i.e., aligning every single modality to image independently. Specifically, ImageBind feeds multi-modal input into corresponding encoders, and adopts for cross-modal contrastive learning. After training on large-scale image-paired data, ImageBind effectively aligns six modalities into a single shared representation space, enabling emergent cross-modal zero-shot capabilities. Based on existing vision-language models, ImageBind can also be utilized for several multi-modal tasks, such as text-to-audio/video retrieval, audio-to-image generation, and audio-referred object detection. Inspired by this, we propose to develop a 3D multi-modal framework that incorporates 3D point clouds with other modalities for general 3D understanding, generation, and instruction following.

3.2. Training Data

To align 3D with multi-modalities, we leverage the pre-trained joint embedding space of ImageBind [22] and utilize contrastive loss [59, 96] to simultaneously align 3D point clouds with the other three modalities: image, text, and audio. To obtain the contrastive training data, we collect a cross-modal dataset of 3D-image-audio-text pairs. There are three steps for dataset collection as follows.

3D-image-text Pairs. We adopt the data pairs of 3D, images, and text from ULIP [84], which includes 3D-image-text triplets built from ShapeNet [8], a common-used dataset containing abundant 3D CAD models. Each 3D point cloud is paired with a corresponding text describing the semantic information of its spatial shape, and a 2D counterpart generated by multi-view image rendering. The text description is constructed by a synset of category names and 64 pre-defined templates.

3D-audio Pairs. To provide more contrastive signals, we also collect the data pairs of 3D and audio from ESC-50 [51] and ShapeNet datasets. Specifically, we first select the categories whose objects can make a sound in the real world from the 55 categories of ShapeNet, such as 'airplane', 'clock', 'washing machine', and 'keyboard'. Then, we preserve only the categories that are also within ESC-50. By this standard, we obtain 9 categories of 3D-audio paired data with extensive audio clips.

3D-image-audio-text Pairs Construction. Finally, we match each 3D-audio pair with its corresponding 3D-image-text data, resulting in a unified 3D-image-audio-text dataset with extensive cross-modal pairs. During training, we simultaneously feed point clouds and their paired data of three modalities for contrastive learning.

3.3. Aligning 3D with Multi-modality

After collecting the 3D paired data, we conduct contrastive training to learn a joint embedding space aligning 3D and multi-modalities. Each data sample contains a point cloud P, along with the paired 2D image I, text description T^s, and audio A, where T^s represents a set of 64 pre-defined templates. For the point cloud, we adopt I2P-MAE [92] as the learnable 3D encoder, denoted as $\mathrm{Encoder_{3D}}(\cdot)$, and append a projection network $\mathrm{Proj}(\cdot)$ of two linear layers, which transforms the encoded 3D feature into ImageBind's multi-modal embedding space. We formulate it as

$$F_{3D} = \mathrm{Proj}(\mathrm{Encoder_{3D}}(P)), \tag{1}$$

where $F_{3D} \in \mathbb{R}^{1\times C}$ denotes the projected 3D embedding, and C equals the feature dimension of ImageBind. For the paired image-text-audio data, we leverage their corresponding encoders from ImageBind for feature extraction, which are frozen during training, formulated as

$$F_{2D}, F_T^s, F_A = \mathrm{ImageBind}(I, T^s, A), \tag{2}$$

where $F_{2D}, F_A \in \mathbb{R}^{1\times C}$ denote the image and audio embeddings, and $F_T^s \in \mathbb{R}^{64\times C}$ denotes the text embedding for a set of 64 descriptions. Then, we conduct an average pooling as

$$F_T = \mathrm{Average}(F_T^s) \quad \in \mathbb{R}^{1\times C}, \tag{3}$$

which represents the aggregated text embedding with more robustness. After that, we adopt contrastive loss [96] between 3D and other modalities, which effectively enforces 3D embeddings to align with the joint representation space, formulated as

$$L_{total} = L(F_{3D}, F_{2D}) + L(F_{3D}, F_T) + L(F_{3D}, F_A).$$

Note that some categories of the training data do not include the paired audio A, since they inherently cannot make any sound, e.g., bottle, planter, and couch, for which we ignore their audio features and loss.

3.4. Multi-modal Applications

Benefiting from the joint embedding space of Point-Bind, we respectively introduce several emergent application scenarios concerning 3D and multi-modalities.

Any-to-3D Generation. Inherited from 2D generative models, existing 3D generation methods can only achieve text-to-3D synthesis. In contrast, with the joint embedding space of Point-Bind, we can generate 3D shapes conditioned on any modalities, i.e., text/image/audio/point-to-mesh. In detail, we directly connect the multi-modal encoders of Point-Bind with the pre-trained decoders of current CLIP-based text-to-3D models, e.g., CLIP-Forge [65]. Without further training, we can synthesize a 3D car mesh based on an input car horn.

3D Embedding-space Arithmetic. We observe that 3D features encoded by Point-Bind can be directly added with other modalities to incorporate their semantics, further achieving composed cross-modal retrieval. For instance, the combined embeddings of a 3D car and audio of sea waves can retrieve an image showing a car parking by a beach, while the composition of a 3D laptop and audio of keyboard typing can retrieve an image of someone who is working with a laptop.

3D Zero-shot Understanding. For traditional text-inferred 3D zero-shot classification, Point-Bind attains *state-of-the-art* performance guided by additional multi-modal supervision. Besides, Point-Bind can also achieve audio-referred 3D open-world understanding, i.e., recognizing 3D shapes of novel categories indicated by the corresponding audio data [51].

4. Point-LLM

In this section, we illustrate how to leverage Point-Bind to develop 3D large language models (LLMs), termed as Point-LLM, which fine-tunes LLaMA [73] to achieve 3D question answering and multi-modal reasoning. The overall pipeline of Point-LLM is shown in Figure 5.

4.1. 3D Instruction-following Capacity

Our Point-LLM is developed on top of ImageBind-LLM [19], which conducts multi-modality instruction tuning by injecting the semantics of ImageBind into LLaMA. Our approach exhibits both data and parameter efficiency.

No Need for 3D Instruction Data. During training, only the public vision-language data [9, 13, 68, 70] is required for fine-tuning LLaMA to learn the 3D-conditioned response capacity. As Point-Bind has built a joint embedding space between 3D and multi-modalities, if any one of the modalities is trained to connect with LLaMA, the others would also be aligned at the same time. Considering this, we select the 2D image modality, since it has the most public data with paired language. By only aligning ImageBind's image encoder with LLaMA, we avoid the expensive cost of collecting and annotating large-scale 3D instruction data, thus saving extensive resources.

Parameter-efficient Training. Instead of tuning the entire LLMs [38, 97], we only unfreeze partial parameters within LLaMA for efficient vision-language instruction tuning. Specifically, a learnable bind network is adopted to bridge the image encoder of ImageBind with the language space of LLaMA. Then, a zero-initialized gating mechanism is proposed to add the image features after the bind

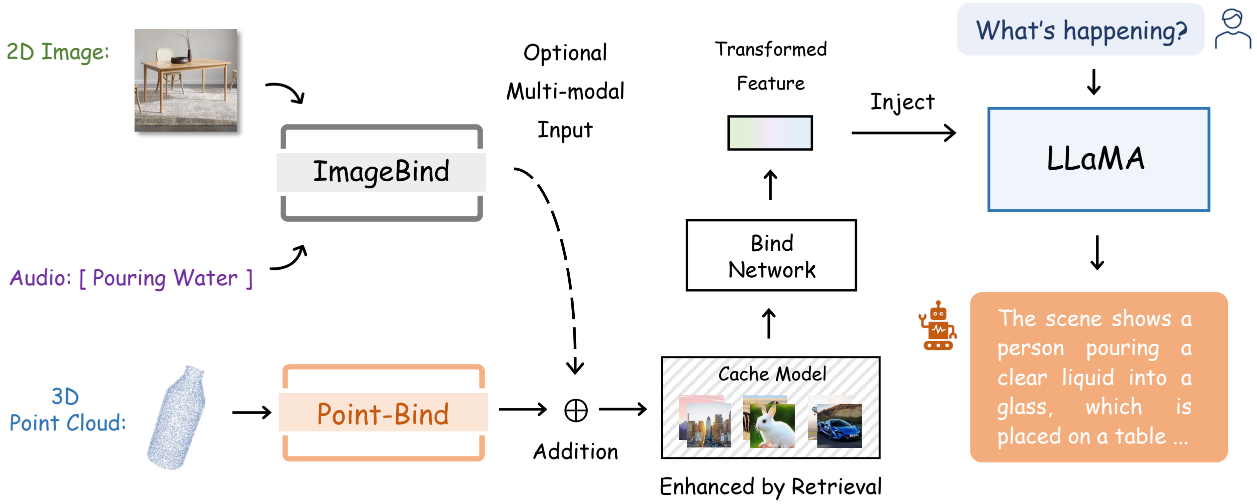


Figure 5. **Inference Paradigm of Point-LLM.** Referring to ImageBind-LLM [19], we adopt a bind network, a visual cache model, and zero-initialized gating mechanisms to fine-tune LLaMA [73] to follow 3D instructions. Optionally, our Point-LLM can also take as input multi-modality data, and conduct cross-modal reasoning for language response.

network to the words tokens within LLaMA. This mechanism can progressively inject visual instruction cues into LLaMA for stable training at early stages, inspired by LLaMA-Adapter [91]. By such a parameter-efficient fine-tuning strategy, most parameters of LLaMA are kept frozen, and only the zero-initialized gating factors and bias-norm weights [91] are learnable for fine-tuning. Please refer to ImageBind-LLM [19] for further training details. After the vision-language training, the joint embedding space enables LLaMA to naturally align with other modalities, such as audio within ImageBind and 3D point clouds of Point-Bind. Therefore, our Point-LLM effectively provides LLaMA with 3D-instruction following capacity without any 3D instruction data, indicating superior data efficiency.

4.2. 3D Question Answering

For an input language instruction and a 3D point cloud, we feed them into the fine-tuned LLaMA and our Point-Bind, respectively. Then, the encoded 3D feature is enhanced by a visual cache model proposed in ImageBind-LLM, before feeding into the bind network. The cache model is only adopted during inference, and constructed in a training-free manner [94].

Enhancement by Visual Cache. As we adopt the image encoder of ImageBind for training, but switch to Point-Bind's 3D encoder for inference, the cache model is designed to alleviate such modality discrepancy for better 3D geometry understanding. Referring to ImageBind-LLM, the cache model stores from three ImageBind-encoded image features from the training data, which are regarded as both keys and values for knowledge retrieval. We regard the input 3D feature as the query, and retrieve the top-k similar visual keys from the cache model. Then, according to the cosine similarity, we aggregate the corresponding cached values (top-k similar image features), and add the result to the original 3D feature via a residual connection. In this way, the enhanced 3D feature can adaptively incorporate similar visual semantics from the cache model. This boosts the representation quality of 3D shapes, and mitigates the semantic gap of 2D-3D encoders within Point-LLM. After this, the enhanced feature is fed into the bind network for feature transformation and LLaMA for response generation.

3D and Multi-modal Reasoning. In addition to point clouds, our Point-LLM can also conduct cross-modal reasoning and generate responses conditioned on multiple modalities. For an additional input image or audio, we utilize the image or audio encoder of ImageBind to extract the features, and directly add them with the 3D feature encoded by Point-Bind. By injecting such integrated features into LLaMA, Point-LLM can reason cross-modal semantics, and respond with the information of all input modalities. This demonstrates the promising significance of aligning multi-modality with 3D LLMs.

5. Experiments

In this section, we first present the implementation details of the multi-modal training for Point-Bind. Then, we illustrate the emergent multi-modal applications, i.e., Point-LLM for 3D instruction following, 3D cross-modal retrieval, 3D embedding-space arithmetic, any-to-3D gener-

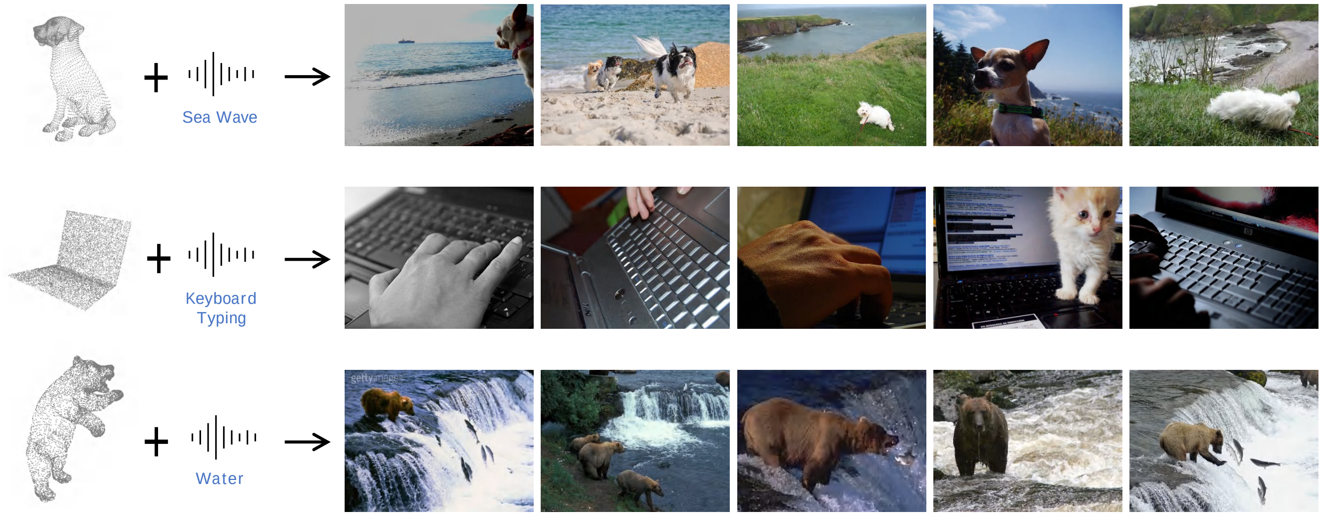


Figure 6. **Embedding-space Arithmetic of 3D and Audio.** We demonstrate Point-Bind's capability for multi-modal semantic composition by retrieving 2D images with a combination of 3D point cloud and audio embeddings.

ation, and 3D zero-shot understanding. Finally, we conduct an ablation study to verify the effectiveness of our designs.

5.1. Implementation Details

We adopt pre-trained I2P-MAE [92] as the 3D encoder of Point-Bind, and utilize the collected 3D-image-text-audio pairs for pre-training. We only update the 3D encoder with the newly added projection network, and freeze the encoders of other modalities in ImageBind [22]. The projection network is simply composed of two linear layers with an intermediate LayerNorm [4]. We train Point-Bind for 300 epochs with a batch size of 64, and adopt AdamW [44] as the optimizer with a learning rate of 0.003.

5.2. Point-LLM for 3D Q&A

Settings. We refer to ImageBind-LLM [19] to conduct parameter- and data-efficient fine-tuning to inject 3D instructions into the pre-trained LLaMA 7B model [73]. In detail, the fine-tuning techniques include zero-initialized gating [20, 91], LoRA [29], and bias-norm tuning [18, 21, 81, 88]. We utilize a collection of several datasets [9, 68, 70] for vision-language training, and require no 3D instruction-following dataset due to the learned joint embedding space.

Analysis. In Figure 3, we provide the question-answering examples of Point-LLM, which shows favorable 3D instruction-following and multi-modal reasoning capacity. As shown, for either English or Chinese instructions, Point-LLM can effectively incorporate the spatial geometry of input point clouds and generate detailed language responses. It obtains a comprehensive 3D understanding for both global and local characteristics, e.g., recognizing the pattern of the piano keyboard and the shape of the airplane's wing and tail. Then, our Point-LLM can also respond with cross-modal understanding. For an input 3D model with a 2D image or audio, Point-LLM can enable LLaMA to take both two conditions into understanding and reasoning, which thus incorporates multi-modal semantics in the output language response. With superior data- and parameter-efficiency, the examples indicate the 3D multi-modal instruction-following capabilities of Point-LLM.

Table 1. **Performance on 3D Cross-modal Retrieval**, including 3D-to-3D, 2D-to-3D, 3D-to-2D, and text-to-3D retrieval. We report the mAP scores (%) on ModelNet40 [80] dataset.

Method	3D → 3D	2D → 3D	3D → 2D	Text → 3D
PointCLIP [90]	37.63	13.12	5.28	10.86
PointCLIP-V2 [100]	47.94	20.48	9.22	52.73
ULIP [84]	60.58	20.30	29.75	50.51
Point-Bind	**63.23**	**34.59**	**42.83**	**64.50**
Gain	+2.65	+14.29	+13.08	+13.99

5.3. 3D Cross-modal Retrieval

To evaluate the multi-modal alignment of Point-Bind, we experiment on several cross-modal retrieval tasks, i.e., 3D-to-3D, 2D-to-3D, 3D-to-2D, and text-to-3D retrieval.

Settings. We conduct 3D zero-shot retrieval on multi-modal ModelNet40 [80] dataset, which contains 9,843 CAD models for training and 2,468 for testing of 40 categories. ModelNet40 provides data of three modalities for retrieval, i.e., image, point cloud, and mesh. We obtain the retrieved results by ranking the similarities between embed-

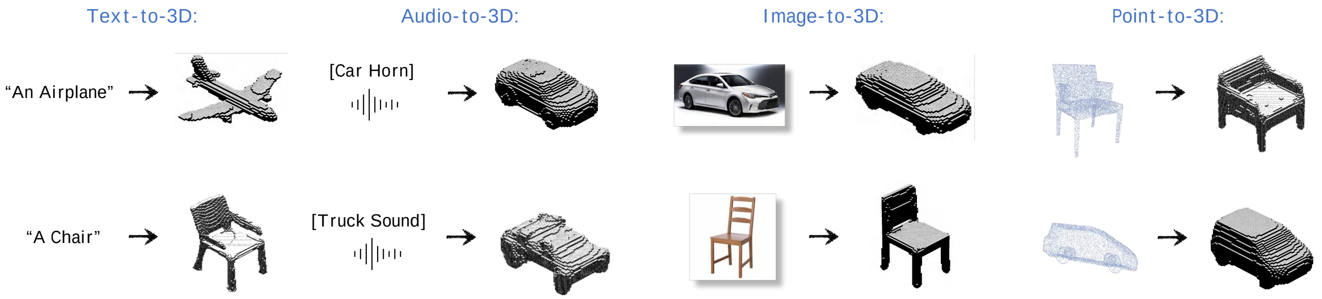


Figure 7. **Any-to-3D Generation.** Besed on CLIP-Forge [65], our constructed joint embedding space can effectively generate 3D mesh models conditioned on text, audio, image, and point cloud input.

Table 2. **Performance of 3D Zero-shot Classification.** We report the classification accuracy (%) on ModelNet40 [80].

Model	Encoder	Performance
PointCLIP [90]	CLIP	20.2
ULIP [84]	Point-BERT	60.4
PointCLIP V2 [100]	CLIP	64.2
ULIP 2 [85]	Point-BERT	66.4
Point-Bind	Point-BERT	76.3
Point-Bind	**I2P-MAE**	**78.0**
Gain	-	+11.6

dings of point clouds and other modalities. Following previous works [33,42], we measure the networks via the mean Average Precision (mAP) score, a commonly used evaluation criterion for retrieval tasks.

Analysis. In Table 1, we report the quantitive results for 3D zero-shot retrieval, where Point-Bind attains *state-of-the-art* performance on all benchmarks compared with prior works. In particular, for 2D-to-3D and text-to-3D retrieval, Point-Bind surpasses the second-top ULIP [84] significantly by **+14.29%** and **+13.99%** improvements, respectively. This indicates the superior cross-modal understanding capacity of our approach.

5.4. Embedding-space Arithmetic with 3D

With the multi-modal alignment, we further explore the capability of embedding composition, i.e., the embedding-space arithmetic of 3D and other modalities, e.g., audio.

Settings. To obtain the multi-modal input for arithmetic, we utilize 3D objects from ShapeNet [8] and TextANIMAR2023 [7], and audio clips from ESC-50 [51]. We simply add the 3D and audio embeddings from Point-Bind and ImageBind, respectively, and then retrieve 2D images from ImageNet [14] with 1,000 image categories.

Analysis. In Figure 6, we show the results of 2D image retrieval with the composed embeddings between 3D and audio. As shown in the first row, with the combined embeddings of a 3D dog and sea-wave audio, we effectively retrieve 2D images of dogs by the sea. Similarly, with the combination of a 3D laptop and keyboard-typing audio, the obtained images show someone is working with a laptop, or a cat inadvertently presses on the keyboard. Likewise, the last row retrieves images of bears hunting by the water by using embeddings of a 3D bear and audio of flowing water. The examples demonstrate that the 3D features encoded by Point-Bind can be directly added with other modalities, and well incorporate their semantics, achieving favorable composed cross-modal retrieval capacity.

5.5. Any-to-3D Generation

Settings. Existing text-to-3D generation methods normally adopt CLIP's text encoder to process the input language prompt. Considering this, we simply replace it with the multi-modalities encoders of Point-Bind and ImageBind without further training, which follows the original generative decoder for 3D shape synthesis. We adopt the decoder of CLIP-Forge [65] by default.

Analysis. In Figure 7, we show the examples of any-to-3D generation powered by Point-Bind. For text, audio, and point cloud prompts, our approach can all produce satisfactory 3D meshes. This demonstrates the well-aligned embedding space of 3D and multiple modalities.

5.6. 3D Zero-shot Understanding

In this section, we test the open-word understanding ability of Point-Bind, i.e., recognizing novel classes, by 3D zero-shot classification on ModelNet40 [80] dataset.

Settings. Following previous works, we utilize the text embeddings from CLIP's [59] or ImageBind [22]'s text encoder to construct the zero-shot classification head. Specifi-

Table 3. **Ablation Study** exploring different designs of the projection network and 3D encoders. We report the results (%) for zero-shot classification on ModelNet40 [80].

Projection	Acc.	3D Encoder	Acc.
One Linear	76.46	PointNeXt [57]	67.96
Two Linear	**78.00**	Point-BERT [86]	76.70
Three Linear	76.78	I2P-MAE [92]	**78.00**

cally, we apply a simple template of *'a/an [CLASS]'* for the 40 categories of ModelNet40, and calculate the cosine similarity between 3D and all textual embeddings, selecting the most similar one as the final prediction.

Analysis. We report the 3D zero-shot classification accuracy in Table 2, where our Point-Bind surpasses existing methods with *state-of-the-art* performance. This indicates the unified representation space of Point-Bind leads to strong emergent 3D open-world recognition.

5.7. Ablation Study

To investigate the effectiveness of our designs in Point-Bind, we conduct ablation studies on the projection network and 3D encoders in Table 3. We report the performance of zero-shot classification on ModelNet40 [80] dataset. In the first two columns, we experiment with different projection schemes for embeddings after the 3D encoder. As shown, using two linear layers for embedding projection performs the best. In the last two columns, we utilize different 3D encoders in Point-Bind, i.e., Point-BERT [86], PointNeXt [57], and I2P-MAE [92]. As reported, the self-supervised Point-BERT and I2P-MAE achieve much better performance, indicating the importance of 3D pre-training to boost the multi-modal alignment.

6. Conclusion

In this paper, we propose **Point-Bind**, a 3D multi-modality model that aligns 3D point clouds with multi-modalities, guided by ImageBind. By aligning 3D objects with their corresponding image-audio-text pairs, Point-Bind obtains a joint embedding space, and exhibits promising 3D multi-modal tasks, such as any-to-3D generation, 3D embedding arithmetic, and 3D open-world understanding. Upon that, we further introduce **Point-LLM**, the first 3D large language model (LLM) with instruction-following capability in both English and Chinese. Our future work will focus on aligning multi-modality with more diverse 3D data, such as indoor and outdoor scenes, which allows for wider application scenarios.

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. CrossPoint: Self-supervised Cross-modal Contrastive Learning for 3D Point Cloud Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 1, 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 1

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19129–19139, 2022. 1

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 8

[5] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 4

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[7] TextANIMAR Challenge. TextANIMAR. https://aichallenge.hcmus.edu.vn/textanimar, 2023. 9

[8] 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. 5, 9

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3, 6, 8

[10] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. PiMAE: Point Cloud and Image Interactive Masked Autoencoders for 3D Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5291–5301, 2023. 4

[11] Siheng Chen, Baoan Liu, Chen Feng, Carlos Vallespi-Gonzalez, and Carl Wellington. 3D Point Cloud Processing and Learning for Autonomous Driving: Impacting Map

Creation, Localization, and Perception. *IEEE Signal Processing Magazine*, 38(1):68–86, 2020. 1

[12] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023. 1

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 6

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 9

[15] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 4

[16] 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. 4

[17] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv preprint arXiv:2106.11097*, 2021. 4

[18] Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020. 8

[19] Peng Gao, Jiaming Han, Chris Liu, Ziyi Lin, Renrui Zhang, and Ziyu Guo. ImageBind-LLM. https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/imagebind_LLM, 2023. 3, 6, 7, 8

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*, 2023. 3, 8

[21] Angeliki Giannou, Shashank Rajput, and Dimitris Papailiopoulos. The expressive power of tuning only the norm layers. *arXiv preprint arXiv:2302.07937*, 2023. 8

[22] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space to Bind Them All. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2, 4, 5, 8, 9

[23] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 1, 4

[24] Ziyu Guo, Yiwen Tang, Renrui Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding with gpt and prototype guidance. *arXiv preprint arXiv:2303.16894*, 2023. 1, 4

[25] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 746–754, 2023. 4

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *arXiv:2111.06377*, 2021. 4

[27] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognition. *arXiv preprint arXiv:2303.11313*, 2023. 4

[28] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 3D Concept Learning and Reasoning From Multi-View Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023. 1

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[30] 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. *arXiv preprint arXiv:2210.01055*, 2022. 4

[31] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[32] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-Shot Text-Guided Object Generation with Dream Fields. 2022. 4

[33] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal Center Loss for 3D Cross-modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021. 9

[34] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. PU-GAN: a point cloud upsampling adversarial network. In *International Conference on Computer Vision*, 2019. 4

[35] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1

[36] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7005–7015, 2021. 4

[37] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring Cross-Video and Cross-Modality Signals for Weakly-Supervised Audio-Visual Video Parsing. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6

[39] 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. *arXiv preprint arXiv:2305.10764*, 2023. 1, 4

[40] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2D: Contrastive Pixel-to-point Knowledge Transfer for 3D Pretraining. *arXiv preprint arXiv:2104.04687*, 2021. 4

[41] Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. ISS++: Image as Stepping Stone for Text-Guided 3D Shape Generation. *arXiv preprint arXiv:2303.15181*, 2023. 4

[42] Zhitao Liu, Zengyu Liu, Jiwei Wei, Guan Wang, Zhenjiang Du, Ning Xie, and Heng Tao Shen. Instance-Variant Loss with Gaussian RBF Kernel for 3D Cross-modal Retrieval. *arXiv preprint arXiv:2305.04239*, 2023. 9

[43] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3D Object Detection via Transformers. In *International Conference on Computer Vision*, pages 2949–2958, 2021. 1

[44] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[45] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. *arXiv preprint arXiv:2112.03221*, 2021. 4

[46] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-end Transformer Model for 3D Object Detection. In *International Conference on Computer Vision*, pages 2906–2917, 2021. 1

[47] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. CLIP-Mesh: Generating Textured Meshes from Text using Pretrained Image-text Models. In *ACM SIGGRAPH Asia Conference*, pages 1–8, 2022. 4

[48] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning Audio-Video Modalities from Image Captions. In *European Conference on Computer Vision*, pages 407–426, 2022. 4

[49] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1

[50] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked Autoencoders for Point Cloud Self-supervised Learning. In *European Conference on Computer Vision*, pages 604–621, 2022. 4

[51] Karol J Piczak. ESC: Dataset for Environmental Sound Classification. In *ACM International Conference on Multimedia*, pages 1015–1018, 2015. 5, 6, 9

[52] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[53] Omid Poursaeed, Tianxing Jiang, Quintessa Qiao, Nayun Xu, and Vladimir G. Kim. Self-supervised Learning of Point Clouds via Orientation Estimation. *arXiv preprint arXiv:2008.00305*, 2020. 4

[54] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 4

[55] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv preprint arXiv:1706.02413*, 2017. 4

[56] 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*, 2023. 1, 4

[57] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies. In *Advances in Neural Information Processing Systems*, 2022. 4, 10

[58] Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. Pix4Point: Image Pretrained Transformers for 3D Point Cloud Understanding. *arXiv*, 2022. 4

[59] 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, 2021. 4, 5, 9

[60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[61] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A Multimodal Transformer for Visual Speech Enhancement. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6658–6662, 2021. 4

[62] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-Local Bidirectional Reasoning for Unsupervised Representation Learning of 3D Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[65] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-Forge: Towards Zero-shot Text-to-shape Generation. *arXiv preprint arXiv:2110.02624*, 2021. 4, 6, 9

[66] Jonathan Sauder and Bjarne Sievers. Self-supervised Deep Learning on Point Clouds by Reconstructing Space. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[67] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *International Conference on Computer Vision*, pages 9339–9347, 2019. 1

[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3, 6, 8

[69] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 1

[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018. 3, 6, 8

[71] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 1

[72] Desney S Tan, George G Robertson, and Mary Czerwinski. Exploring 3d navigation: combining speed-coupled flying with orbiting. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 418–425, 2001. 1

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 6, 7, 8

[74] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3D Instance Segmentation on Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 1

[75] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised Point Cloud Pre-Training via Occlusion Completion. In *International Conference on Computer Vision*, pages 9782–9792, 2021. 4

[76] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019. 1

[77] 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. *Advances in Neural Information Processing Systems*, 35:14388–14402, 2022. 4

[78] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019. 1

[79] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023. 1

[80] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. 8, 9, 10

[81] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. *arXiv preprint arXiv:2304.06648*, 2023. 8

[82] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2Point: 3D Point-Cloud Understanding with 2D Image Pretrained Models. In *European Conference on Computer Vision*, pages 638–656, 2022. 4

[83] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3D: Zero-shot Text-to-3D Synthesis using 3D Shape Prior and Text-to-image Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 4

[84] 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. 1, 4, 5, 8, 9

[85] 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. *arXiv preprint arXiv:2305.08275*, 2023. 4, 9

[86] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4, 10

[87] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022. 1

[88] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning

for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 8

[89] 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. 4

[90] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 1, 4, 8, 9

[91] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 7, 8

[92] 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. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 21769–21780, 2023. 1, 2, 4, 6, 8, 10

[93] Renrui Zhang, Liuhui Wang, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *CVPR 2023*, 2023. 4

[94] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022. 7

[95] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 1

[96] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Machine Learning for Healthcare Conference*, pages 2–25, 2022. 5, 6

[97] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 6

[98] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 1

[99] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Hao Dong, and Peng Gao. Less is more: Towards efficient few-shot 3d semantic segmentation via training-free networks. *arXiv preprint arXiv:2308.12961*, 2023. 4

[100] 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. *arXiv preprint arXiv:2211.11682*, 2022. 1, 4, 8, 9