POINT-BIND & POINT-LLM: ALIGNING POINT CLOUD WITH MULTI-MODALITY FOR 3D UNDERSTANDING, GENERATION, AND INSTRUCTION FOLLOWING

Anonymous authors
Paper under double-blind review

A OVERVIEW

- Section B: Additional experiments.
- Section C: Related work.
- Section D: Additional implementation details.

B ADDITIONAL EXPERIMENTS

Cross-modal Retrieval on More Modalities. To verify the potential of Point-Bind to align multi-modalities, we conduct cross-modal retrieval between 3D and more modalities, i.e., video, depth, and infrared data. We utilize the following work of ImageBind (Girdhar et al., 2023), Language-Bind (Zhu et al., 2023a), as guidance, and pre-train Point-Bind under the same paradigm. By aligning 3D with the image space of LanguageBind, Point-Bind achieves a unified space with multi-modalities including video, depth, and infrared data. As shown in Figure 1, with the 3D car/person as input, Point-Bind effectively retrieves corresponding video, depth, and infrared data with the same semantics. This indicates the superior cross-modal understanding capacity of our approach.

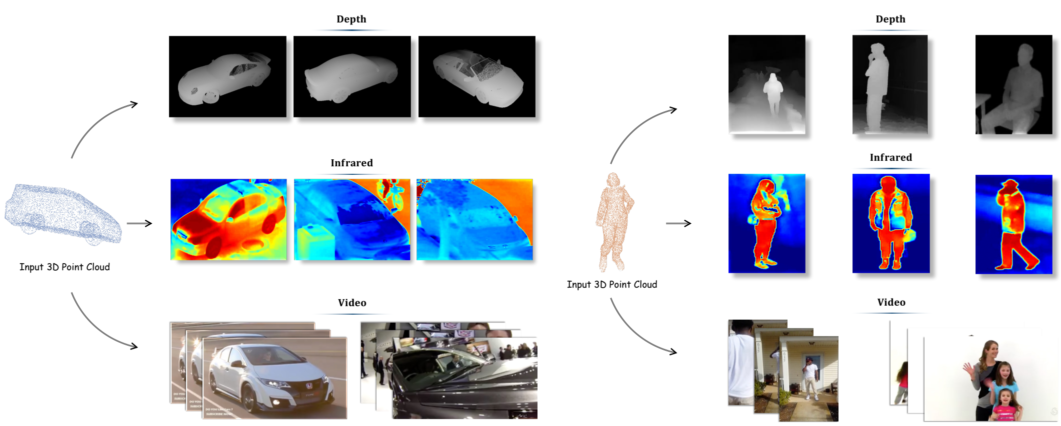

Figure 1: **Additional Visualization of Cross-modal Retrieval.** We visualize the cross-modal retrieval between 3D and three new modalities, i.e., video, depth, and infrared data. Note that, for these modalities, we utilize LanguageBind (Zhu et al., 2023a) as the guidance for pre-training Point-Bind.

Quantitative Results of Any-to-3D Genration. Besides text-to-3D generation, we quantitatively demonstrate the efficacy of Point-Bind on any-to-3D generation in Table 1. We generate the 3D mesh conditioned on multi-modalities and their embedding-space arithmetic, i.e., directly combining embeddings from different modalities to guide 3D generation. We adopt different settings for different modalities. For audio-to-mesh generation, we only generate objects of the car, airplane,

Figure 2: **Any-to-3D Generation** based on CLIP-Forge (Sanghi et al., 2021). Besides ISS (Liu et al., 2022), our Point-Bind is generalized to combine any text-to-3D models for any-to-3D generation.

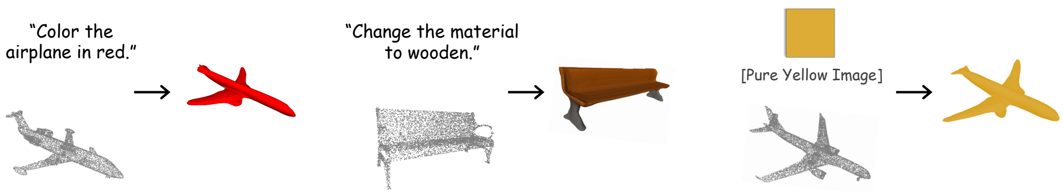

Figure 3: **3D Editing with Multi-modal Instructions.** Within the joint 3D embedding space of Point-Bind, we can effectively edit input 3D point clouds with multi-modal instructions, e.g., language or image.

and boat categories considering the limited class number. We sample 10 audio clips per category from ESC-50 dataset (Piczak, 2015) as input. The airplane take-off sound, car horn, and sea wave sound are selected to generate the airplane, car, and boat categories, respectively. For image-to-mesh generation, we sample 10 images corresponding to ShapNet's 13 categories from ImageNet dataset (Deng et al., 2009) as the 2D prompt. For point-to-mesh synthesis, we sample 10 point clouds per category from the ShapeNet dataset (Chang et al., 2015) as prompt. Compared to text-to-3D generation, the results in Table 1 suggest that Point-Bind can also achieve satisfactory generation quality with other modalities as conditions.

Table 1: **Quantitative Results of Any-to-3D Generation.** We report the Fréchet Inception Distance (FID) and Fréchet Point Distance (FPD) scores for comparison.

Source Modality	FID (↓)	FPD (↓)
Audio	166.97	30.46
Image	95.77	19.41
Point Cloud	86.14	20.13
Image + Text	86.79	26.13
Point Cloud + Text	88.78	26.39
Point Cloud + Image	87.03	21.19

Any-to-3D Generation with CLIP-Forge (Sanghi et al., 2021). Besides ISS (Liu et al., 2022), we also adopt the decoder of CLIP-Forge and show the examples of any-to-3D generation powered by Point-Bind in Figure 2. For text, audio, and point cloud prompts, our approach can all produce satisfactory 3D meshes. This demonstrates that Point-Bind generalizes well and can guide other 3D generation models conditioned on multi-modalities.

3D Editing with Multi-modal Instructions. Besides the any-to-3D generation, our approach can further enable 3D editing with multi-modal instructions, as visualized in Figure 3. For example, given a 3D airplane, we can provide a language instruction, "Color the 3D shape in red", or a pure yellow picture as the visual instruction. Then, we respectively feed them into Point-Bind's 3D encoder and ImageBind's text or image encoder. Due to the joint embedding space, the generative decoder can incorporate their semantics and output the airplane in red/yellow. Likewise, given an ordinary 3D bench, we can provide instructions like "Modify the material to wooden". The model can correspondingly generate a wooden chair. Therefore, benefiting from the emergent capacity of Point-Bind, we can simply achieve any-to-3D generation and editing, exhibiting favorable training efficiency and generalization capability.

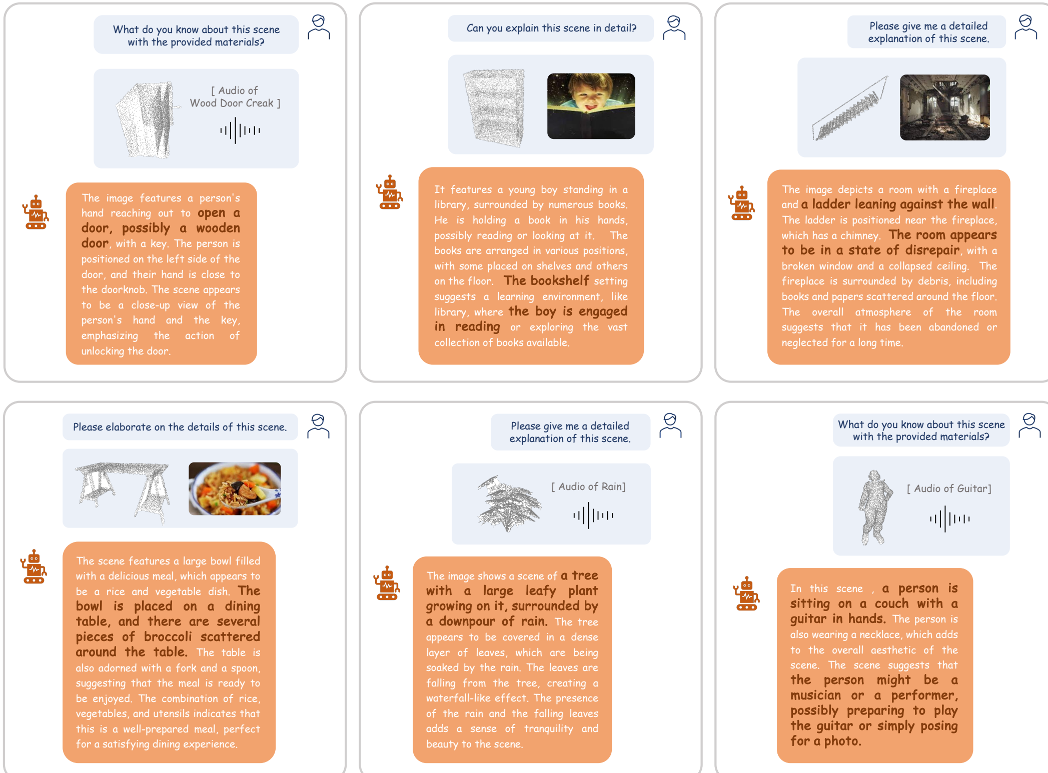

Figure 4: **Additional 3D Question-answering Examples of Point-LLM.** Point-LLM can effectively generate detailed responses and conduct superior cross-modal reasoning, based on the given multi-modal instructions.

Table 2: **Comparison to ULIP by Teacher Models with The Same Image Encoder: ViT-H.**

Method	Teacher Model	Image Encoder	Accuracy
ULIP	OpenCLIP (Ilharco et al., 2021)	ViT-L	60.4%
ULIP	OpenCLIP (Ilharco et al., 2021)	ViT-H	73.2%
Point-Bind	ImageBind (Girdhar et al., 2023)	ViT-H	**76.3%**

Additional Comparison and Analysis with ULIP. The teacher model of Point-Bind, ImageBind (Girdhar et al., 2023), has different pre-training settings with ULIP's (Xue et al., 2022) teacher model, SLIP (Mu et al., 2021). In this paragraph, we compare Point-Bind and ULIP with the same pre-trained teacher models. We first reproduce a ULIP model also pre-trained by CLIP's ViT-H image encoder, which is the same as ImageBind's image encoder. Note that, ImageBind freezes the ViT-H image encoder and text encoder of OpenCLIP during its pre-training. That is, ImageBind and OpenCLIP share the same weights in their image and text encoders. As shown in Table 2, for zero-shot classification on ModelNet40 (Wu et al., 2015), although the ULIP's performance can be improved by the ViT-H image encoder, our approach still performs better via a joint multi-modal embedding space.

Generalizability of Point-Bind with Techniques from JM3D (Wang et al., 2023a). JM3D (Wang et al., 2023a) proposes two delicate approaches to enhance the multi-modal pre-training of 3D models: Structured Multimodal Organizer (SMO) and Joint Multi-modal Alignment (JMA). SMC adopts multi-view rendered images and hierarchical text for more comprehensive representation, and JMA aims to achieve better mult-modal synergy by generating joint vision-language features. We also add the two techniques in JM3D into our Point-Bind for the image and text modal-

Table 3: **Performance(%) of Point-Bind with JM3D (Wang et al., 2023a) and CG3D (Wang et al., 2023a) on 3D Zero-shot Classification and Cross-modal Retrieval Tasks.**

Method	Zero-shot Cls.	3D → 3D	2D → 3D	3D → 2D	Text → 3D
Point-Bind	78.0	63.2	34.6	42.8	64.5
Point-Bind w JM3D	**78.4**	**64.1**	**35.5**	**43.9**	64.7
Point-Bind w CG3D	78.2	63.5	34.3	43.2	**64.8**

ities within ImageBind (Girdhar et al., 2023), and evaluate on two benchmarks: 3D zero-shot classification and cross-modal retrieval on ModelNet40 (Wu et al., 2015). As shown in Table 3, the capabilities of Point-Bind are well enhanced by integrating SMO and JMA, indicating the importance of more comprehensive vision-language guidance.

Generalizability of Point-Bind with Techniques from CG3D (Hegde et al., 2023). CG3D (Hegde et al., 2023) shares a similar contrastive learning paradigm with ULIP, and introduces learnable visual prompts for CLIP's image encoder for better adaption of 2D rendered images. For our Point-Bind, we also add learnable visual prompts to the image encoder of ImageBind, and report the results in Table 3. On both benchmarks, the prompting approach from CG3D can improve the performance of Point-Bind, which demonstrates the effectiveness of fine-tuning the pre-trained image embeddings.

Additional 3D Question-answering Examples. We provide more 3D question-answering examples in Figure 4, showing the 3D instruction-following and multi-modal reasoning capacity of Point-LLM. As shown, given a 3D shape with a 2D image or audio, Point-LLM effectively enables LLaMA (Touvron et al., 2023) injected with multi-modal semantics, and responds with cross-modal understanding and reasoning. Additionally, as shown in Figure 5, we show more examples of Point-LLM for straightforward question answering, e.g., "How to start it?", "What is the purpose of this thing?". Our model can respond with precise answers that correspond to the input point cloud.

Examples of Indoor Scene Understanding. We further implement a scene-level variant of our model, termed Point-$\text{LLM}_{\text{Scene}}$. We focus on the understanding of indoor scenes on ScanNet (Dai et al., 2017), and show the qualitative examples in Figure 6. Specifically, to obtain the scene-level understanding capacity, we fine-tune our object-level Point-LLM by an existing 3D question-answering dataset (Wang et al., 2023c) constructed from ScanRefer (Chen et al., 2020). We add three MLP layers with residual connections between Point-Bind's 3D encoder and the LLM, which is responsible for learning the scene-level 3D geometries. We only enable the new MLP layers to be trainable, while keeping other components frozen to preserve the pre-trained cross-modal knowledge. As shown, our model can respond with detailed and reasonable answers that correspond to the input 3D scene and target object.

C 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 (Desai & Johnson, 2021; Fang et al., 2021; Nagrani et al., 2022), and enhance the cross-modal performance for downstream tasks (Lin et al., 2021b; Ramesh et al., 2021; Botach et al., 2022; Guo et al., 2023c), and video-text-audio integration for text generation (Lin et al., 2021a). The representative vision-language pre-training, CLIP (Radford et al., 2021), effectively bridges the gap between 2D images and texts, which encourages further exploration of cross-modality learning. Recently, ImageBind (Girdhar et al., 2023) 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 (Zhang et al., 2022b; Xue et al., 2022; Afham et al., 2022; Guo et al., 2023a; Chen et al., 2023) 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,

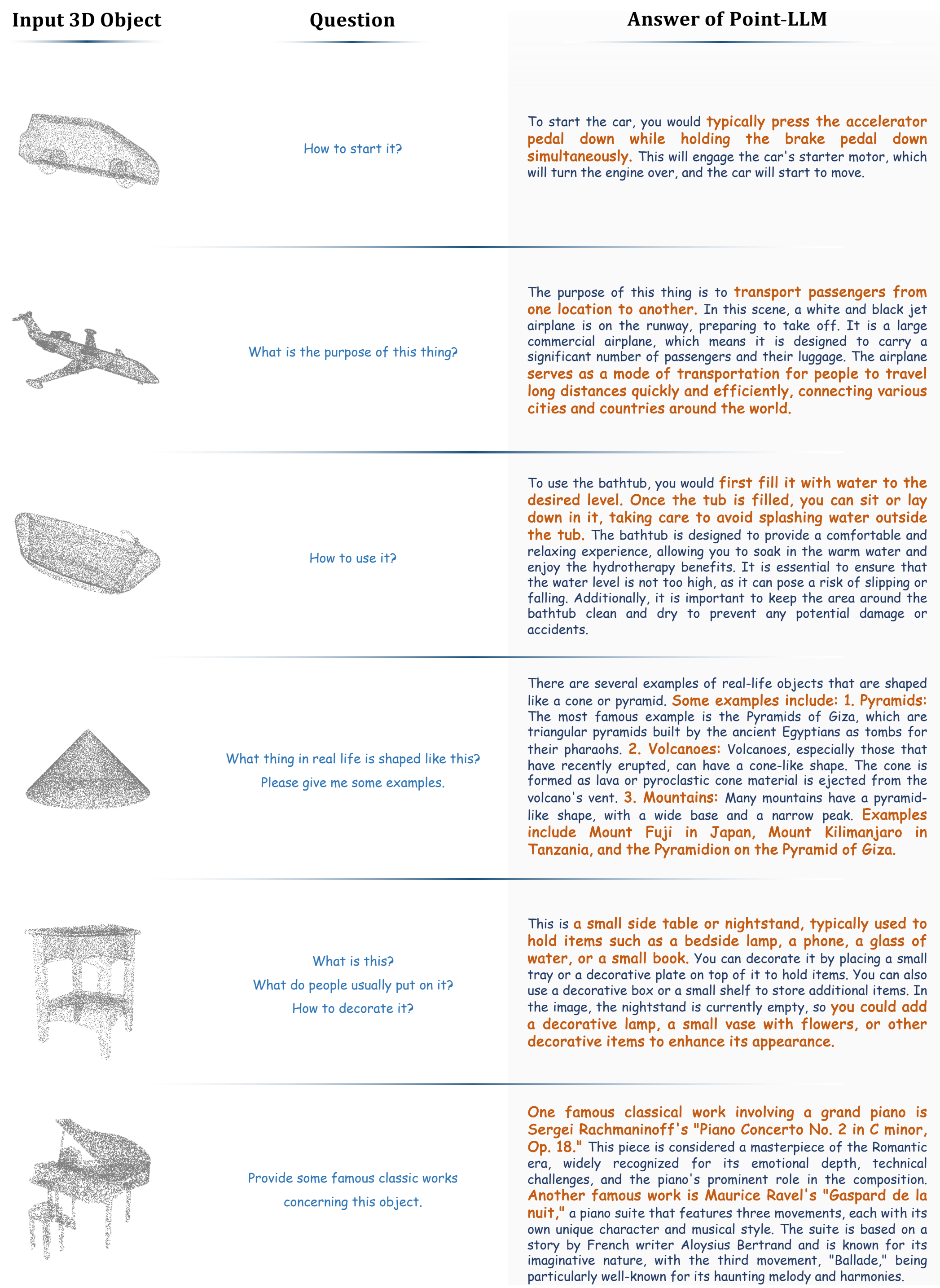

Figure 5: **Examples of Point-LLM for Straightforward Question Answering.**

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

Figure 6: **3D Scene Understanding Capability of Point-LLM$_{\text{Scene}}$.**

2D and language large models, to assist in 3D learning. The prior PointCLIP series (Zhang et al., 2022b; Zhu et al., 2022; Huang et al., 2022) project 3D point clouds into depth maps, and utilize CLIP (Radford et al., 2021) for zero-shot recognition. Image2Point (Xu et al., 2022) instead converts 2D pre-trained models into 3D space as a good network initialization. By contrastive learning, ULIP series (Xue et al., 2022; 2023) and other works (Liu et al., 2023a; Hegde et al., 2023) 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 (Jain et al., 2022a; Sanghi et al., 2021; Xu et al., 2023a; Liu et al., 2023b) or stylized meshes (Mohammad Khalid et al., 2022; Michel et al., 2021) by encoding descriptive textual input. Some works also adopt GPT-3 (Brown et al., 2020) to enhance the language-based understanding of 3D spatial geometry, such as PointCLIP V2 (Zhu et al., 2022) and ViewRefer (Guo et al., 2023b). Different from them, we utilize ImageBind (Girdhar et al., 2023) 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. There are a couple of very recent efforts that introduce LLMs into 3D, concurrent to our Point-LLM. Different from us, they either project 3D data into multi-view images for encoding (Hong et al., 2023), or require large-scale 3D instruction data for fine-tuning (Xu et al., 2023b; Wang et al., 2023b). More importantly, they cannot generate responses conditioned on both 3D and multi-modal input. Thanks to the joint embedding space of Point-Bind, our Point-LLM can discard the expensive 3D instruction tuning, and respond via 3D multi-modal reasoning.

Pre-training in 3D. In recent years, significant progress has been made in supervised learning for 3D vision tasks (Qi et al., 2016; 2017; Qian et al., 2022a; Zhang et al., 2023b; Zhu et al., 2023b). 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 (Chen et al., 2023; Yu et al., 2022; Li et al., 2019; Poursaeed et al., 2020). 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 (Sauder & Sievers, 2019; Wang et al., 2021; Rao et al., 2020). Therein, Point-MAE (Pang et al., 2022) and Point-M2AE (Zhang et al., 2022a) introduce masked autoencoders (He et al., 2021) 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 (Wang et al., 2022; Qian et al., 2022b; Liu et al., 2021; Qi et al., 2023). For example, ACT (Dong et al., 2022) and I2P-MAE (Zhang et al., 2023a) 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.

D Additional Implementation Details

Multi-modal Training of Point-Bind. To align 3D with multi-modalities, we adopt a pre-trained I2P-MAE (Zhang et al., 2023a) as the 3D encoder of Point-Bind by default, and utilize the collected 3D-image-text-audio pairs for pre-training. We utilize a pre-trained ImageBind (Girdhar et al., 2023) with a ViT-H (Dosovitskiy et al., 2020) image encoder. We only update the 3D encoder with the newly added projection network, and freeze the encoders of other modalities in ImageBind. The projection network is composed of two linear layers with an intermediate LayerNorm (Ba et al., 2016). We train Point-Bind for 300 epochs with a batch size of 64, and adopt AdamW (Loshchilov & Hutter, 2017) as the optimizer with a learning rate of 0.003.

3D Cross-modal Retrieval. We utilize ModelNet40 (Wu et al., 2015) to evaluate Point-Bind on cross-modal retrieval tasks without training. The test set of ModelNet40 provides 2,468 samples with two modalities, i.e., 2D images rendered from 3D meshes and corresponding 3D point clouds. We adopt the Mean Average Precision (mAP) score as the criterion, which measures whether the retrieved data belongs to the same class as the query data. We encode 3D point clouds with Point-Bind and conduct four cross-modal retrieval tasks, i.e., 3D-to-3D, 2D-to-3D, 3D-to-2D, and text-to-3D retrieval. For the text prompt, we adopt and separately encoder 64 prompt templates in ULIP (Xue et al., 2022) on each category, and average them as the text embeddings. For the 2D image prompt, we follow (Jing et al., 2021) to utilize multi-view images where the view number is $\in \{1, 2, 4\}$. We average the performance under the three view settings as the final result.

Any-to-3D Generation. We adopt Image as Stepping Stone (ISS) (Liu et al., 2022) to verify Point-Bind's ability of multi-modal feature alignment. We first optimize a projection layer that transfers Point-Bind image features to ISS 3D shape space. Then, we generate 3D shapes from text features based on the pre-trained projection layer and ISS decoder. The ShapeNet (V2) dataset(Chang et al., 2015) with 13 object categories is utilized to train the model. We follow ISS and adopt a text description set with four texts per category. To demonstrate 3D generation quality, we adopt

FID, FPD, and CLIP R-precision as criteria. FID reflects the quality of rendered 2D images from generated 3D shapes. FPD measures the quality of point clouds extracted from generated shapes based on a pre-trained PointNet model (Qi et al., 2016) following ISS. Additionally, we further adopt CLIP R-precision to evaluate the consistency between the text inputs and generated shapes. We build a text description set, which contains our description prompts and 234 additional texts from CLIP-Forge (Sanghi et al., 2021). Then, we perform per-shape CLIP-R-Precision to retrieve the right description for each generated shape and calculate the retrieval accuracy. To give a comprehensive comparison, we mainly compare our approach to three text-to-mesh generation models, CLIP-Forge (Sanghi et al., 2021), Dream Fields (Jain et al., 2022b), and ISS. Note that Dream Fields can not synthesize 3D shapes directly, so we do not need to evaluate its FPD metric. In addition, two baselines, GLIDE/LAFITE+DVR, which first create images and then generate 3D meshes are also included. Following ISS, we first use GLIDE (Nichol et al., 2021) or LAFITE (Zhou et al., 2022) to create 2D images and then generate 3D shapes via DVR (Niemeyer et al., 2020).

REFERENCES

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*, pp. 9902–9912, 2022. 4

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 7

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*, pp. 4985–4995, 2022. 4

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 6

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, 7

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*, pp. 5291–5301, 2023. 4, 7

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 202–221. Springer, 2020. 4

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017. 4

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*, pp. 248–255, 2009. 2

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

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

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*, pp. 15180–15190, 2023. 1, 3, 4, 7

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*, 2023a. 4

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*, 2023b. 7

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, pp. 746–754, 2023c. 4

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *arXiv:2111.06377*, 2021. 7

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, 6

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *arXiv*, 2023. 7

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. 6

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below. 3

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-Shot Text-Guided Object Generation with Dream Fields. 2022a. 6

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 867–876, 2022b. 8

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*, pp. 3142–3151, 2021. 7

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. 7

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*, pp. 7005–7015, 2021a. 4

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, 2021b. 4

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*, 2023a. 6

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. 7

Zhengzhe Liu, Peng Dai, Ruihui Li, XIAOJUAN QI, and Chi-Wing Fu. Iss: Image as stepping stone for text-guided 3d shape generation. In *International Conference on Learning Representations*, 2022. 2, 7

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*, 2023b. 6

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

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. 6

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*, pp. 1–8, 2022. 6

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 3

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*, pp. 407–426, 2022. 4

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 8

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3504–3515, 2020. 8

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*, pp. 604–621, 2022. 7

Karol J Piczak. ESC: Dataset for Environmental Sound Classification. In *ACM International Conference on Multimedia*, pp. 1015–1018, 2015. 2

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. 7

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. 7, 8

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. 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*, 2023. 7

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*, 2022a. 7

Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. Pix4Point: Image Pretrained Transformers for 3D Point Cloud Understanding. *arXiv*, 2022b. 7

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*, pp. 8748–8763, 2021. 4, 6

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*, pp. 6658–6662, 2021. 4

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. 7

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. 2, 6, 8

Jonathan Sauder and Bjarne Sievers. Self-supervised Deep Learning on Point Clouds by Reconstructing Space. *Advances in Neural Information Processing Systems*, 32, 2019. 7

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. 4

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*, pp. 9782–9792, 2021. 7

Haowei Wang, Jiji Tang, Jiayi Ji, Xiaoshuai Sun, Rongsheng Zhang, Yiwei Ma, Minda Zhao, Lincheng Li, Zeng Zhao, Tangjie Lv, et al. Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3403–3414, 2023a. 3, 4

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023b. 7

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes, 2023c. 4

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. 7

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*, pp. 1912–1920, 2015. 3, 4, 7

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*, pp. 638–656, 2022. 6

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*, pp. 20908–20918, 2023a. 6

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023b. 7

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. 3, 4, 6, 7

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. 6

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. 7

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*, 2022a. 7

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*, pp. 8552–8562, 2022b. 4, 6

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*, pp. 21769–21780, 2023a. 7

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*, 2023b. 7

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17907–17917, 2022. 8

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023a. 1

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. 6, 7

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*, 2023b. 7