

Supplementary Materials:

MiniGPT-3D: Efficiently Aligning 3D Point Clouds with Large Language Models using 2D Priors

Anonymous Authors

A QUALITATIVE RESULTS

We present more qualitative results of our MiniGPT-3D, encompassing 3D recognition and captioning, 3D question answering, as well as qualitative comparisons.

A.1 3D Recognition and Captioning

Figure 1 and Figure 2 further showcase the qualitative results of our MiniGPT-3D in 3D recognition and captioning. Given a 3D point cloud and instruction, MiniGPT-3D is capable of generating text responses that include the object’s category, quantity, color, and as well as unique characteristics. Furthermore, our MiniGPT-3D also leverages the point cloud information to make reasonable reasoning, deducing potential uses and emergence timelines. This excellent comprehension of point clouds underscores the advantage of employing priors from 2D-LLMs to build 3D-LLMs.

A.2 3D Question Answering

Figure 3 further provides the qualitative results of our MiniGPT-3D on 3D question answering. Our MiniGPT-3D supports multi-turn dialogues with users regarding the input 3D point cloud. Users can continuously pose various open-ended questions to MiniGPT-3D about the 3D object, such as its working principle, the number of objects, specific historical event times, and even logical questions. Despite only training 47.8M trainable parameters on one single NVIDIA RTX 3090 GPU for 27 hours, through these examples, we observe that our MiniGPT-3D possesses extensive general knowledge and maintains contextual coherence in multi-turn dialogues, outputting correct text responses. These impressive results underscore the superiority of efficiently aligning 3D point clouds with LLMs based on 2D-LLM knowledge.

A.3 Qualitative comparisons

We present more qualitative comparisons, similar to Table 4 in our main paper. The results are shown in Table 1. Compared with other methods, our MiniGPT-3D outputs a more detailed text response, while accurately recognizing object categories and capturing more 3D point cloud information, such as usage, shape, internal components, geometric attributes, materials, etc. The results show the excellent point cloud understanding capabilities of MiniGPT-3D.

B TRAINING DETAILS

This section presents the training details of MiniGPT-3D, encompassing the training settings, model parameter, and the variation in loss across the four training stages.

Training Settings. Table 2 shows the detailed training settings for MiniGPT-3D. Specifically, we use the point-text instruction dataset [7] as the training dataset, encompassing 660k brief captions

and 70k detailed captions & conversations. Within this setup, stages I and II employ the brief captions as their training dataset, while detailed captions & conversations are utilized in stages III and IV. Notably, stages III and IV utilize different types of training data from detailed captions & conversations based on a specific sampling ratio. For optimization, we adopt the AdamW optimizer with a weight decay of 0.05 and employ a cosine decay with a linear warm-up learning rate schedule. The initial learning rate gradually decreases as the training stage progresses.

Regarding the hyperparameters of model components, the point cloud encoder is configured consistently with Point-BERT [8], receiving point cloud data inputs of 8192 points. The point cloud projection layer consists of a two-layer MLP network that transforms the 384-dimensional features output from the point cloud encoder to the input dimension of 1408 for the Value and Key layers in Q-Former [4]. Our proposed Mixture of Query Experts (MQE) comprises eight query experts and an expert router. The expert router includes a two-layer MLP network and a softmax operation, outputting the probability distribution for activating the eight query experts. We activate the two experts with the highest probabilities in our experiments. Q-Former consists of 12 blocks, with each attention module containing 12 attention heads. LoRA [3] is used for efficiently fine-tuning the Q-Former, where the rank and alpha of LoRA are set to 8 and 16, respectively. The modality projector consists of a two-layer MLP that transforms the 768-dimensional point cloud queries output from Q-Former to 2560-dimensional point tokens. The large language model backbone comprises 32 blocks. We efficiently fine-tune the LLM using LoRA, with the rank and alpha of LoRA set to 64 and 16, respectively.

Model Parameter. MiniGPT-3D boasts a total of 2.95 B model parameters, yet we only train 47.8 M parameters on a single RTX 3090 (24G) GPU, which took 27 hours. The specific trainable and frozen model modules are detailed in Figure 4a and Table 4b.

Training Loss. Figure 5 shows the changes in loss across the four training stages of MiniGPT-3D. The scale interval on the horizontal axis corresponds to the duration of training. During stage I, though training only point cloud projection layer (MLP), we observe a steady decrease in loss. During stage II, more modules are fine-tuned on the same dataset as stage I, enhancing the model’s learning capacity, and leading to a continued decrease in loss from the end of stage I. During stage III, the introduction of more challenging tasks temporarily increases the loss compared to the end of stage II, followed by a gradual reduction. During stage IV, only MQE is trained. Since the expert router of MQE is trained from scratch, the loss suddenly increases compared to the end of stage III, but then gradually decreases to the same level or even lower.

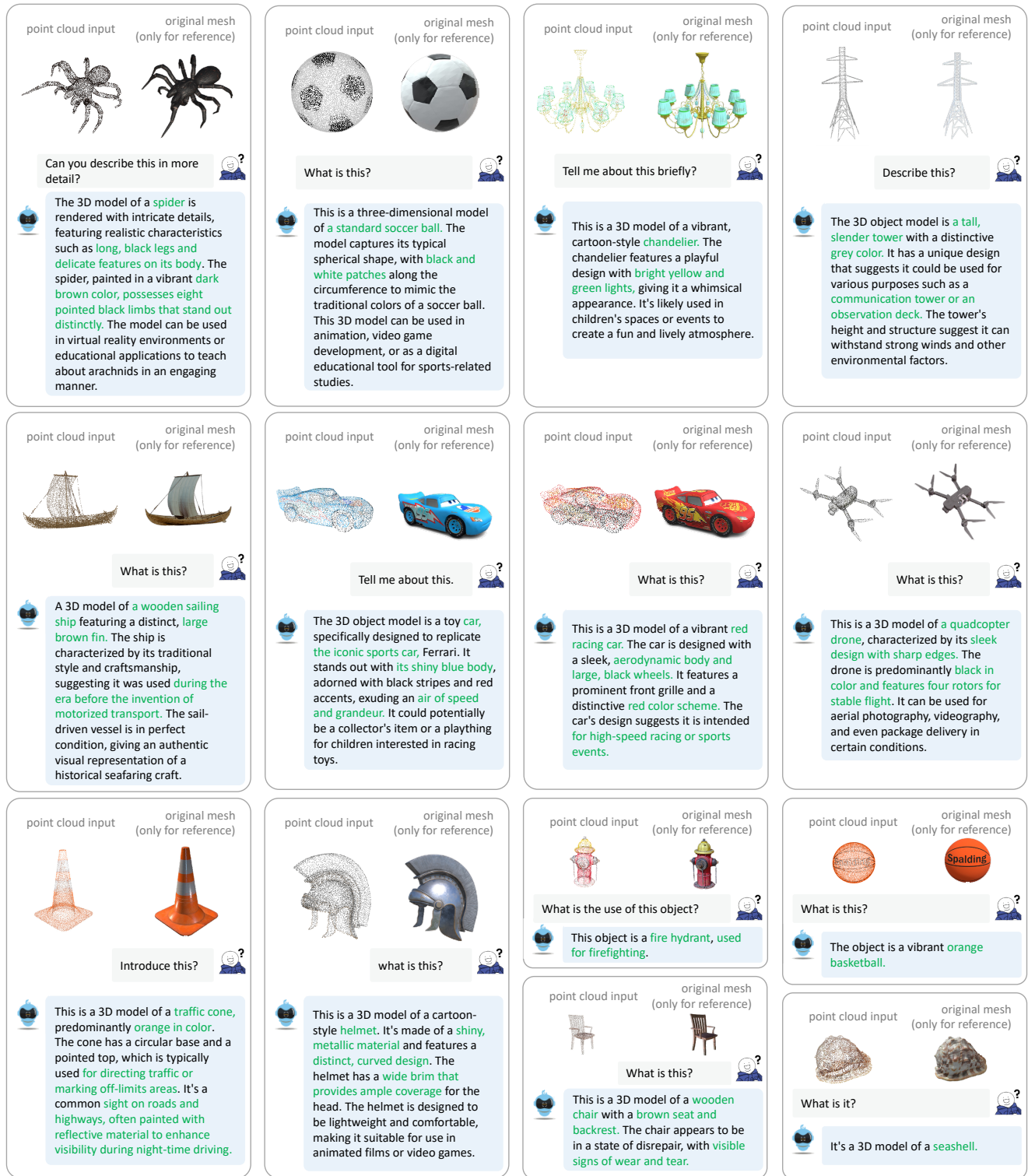


Figure 1: 3D recognition and captioning results (#1) of our MiniGPT-3D. Given a 3D point cloud and a corresponding question, MiniGPT-3D generates a text response. The correct and fine-grained answers are highlight in green.

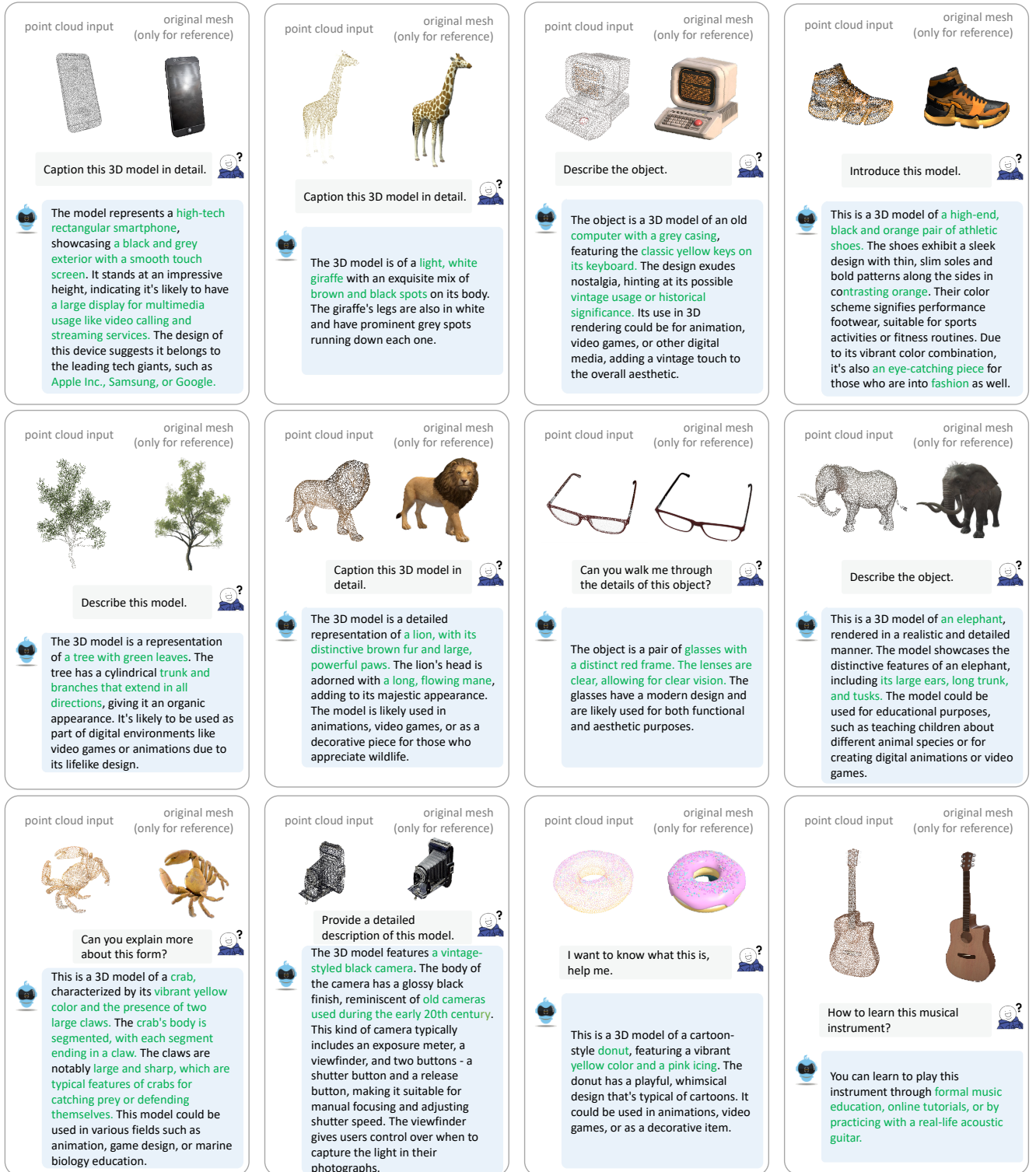


Figure 2: 3D recognition and captioning results (#2) of our MiniGPT-3D. Given a 3D point cloud and a corresponding question, MiniGPT-3D generates a text response. The correct and fine-grained answers are highlight in green.







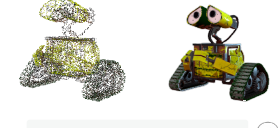

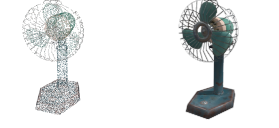
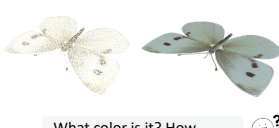


<p>point cloud input original mesh (only for reference)</p>  <p>Why does it have to work at night?</p> <p>It's designed to work at night because it has a light bulb.</p> <p>How many light bulbs are there?</p> <p>There are two light bulbs in this lamp.</p> <p>Why doesn't it light up during the day when there is sunlight?</p> <p>The lamp is designed to be used during the night when there's no sunlight.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>Why do we need it?</p> <p>A watch is a timepiece that helps us keep track of time.</p> <p>Why can it move automatically?</p> <p>The watch has a spring mechanism that allows it to move automatically.</p> <p>Why are there three pointers?</p> <p>The three pointers on the watch are the hour hand, minute hand, and second hand.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What's the use of it?</p> <p>The helicopter is used for various purposes such as transportation, rescue missions, and military operations.</p> <p>What is the function of a helicopter's tail rotor?</p> <p>The tail rotor of a helicopter is used to control the direction and speed of the aircraft.</p> <p>Can it take off without a tail rotor?</p> <p>No, a helicopter cannot take off without a tail rotor. The tail rotor is essential for maintaining balance and stability during flight.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What is the source of piano sound?</p> <p>The source of piano sound is a set of strings.</p> <p>Detailed explanation of the principle of sound production.</p> <p>The piano produces sound through the vibration of its strings. When a key is pressed, it causes a small hammer to strike a string, causing it to vibrate and produce sound.</p> <p>Does it need to use electricity to make a sound?</p> <p>No, the piano doesn't need electricity to make a sound. It uses the kinetic energy from the striking of the keys.</p>
<p>point cloud input original mesh (only for reference)</p>  <p>What does this animal usually eat?</p> <p>This animal, a rabbit, is known to eat plants.</p> <p>How many years can rabbits generally live?</p> <p>Rabbits can live up to 10 years or more, depending on their breed and care.</p> <p>Do rabbits drink water?</p> <p>Yes, rabbits drink water. They need to stay hydrated, especially in hot weather.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What is it?</p> <p>It's a black vintage telephone.</p> <p>Does everyone still use this style now?</p> <p>No, it is more commonly found in vintage collections or nostalgic settings.</p> <p>Who invented the first Telephone?</p> <p>The first telephone was invented by Alexander Graham Bell in 1876.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What can it be used for?</p> <p>It can be used for automated cleaning tasks, such as house cleaning, industrial cleaning, or any other scenarios that require consistent, thorough cleaning.</p> <p>Its tires are very strange, why is it designed like this?</p> <p>The tires' unusual design serves a functional purpose, aiding in the robot's mobility and traction on different surfaces.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What is the function of radio antenna?</p> <p>The radio antenna is used to pick up radio signals from a broadcasting station.</p> <p>Who invented the first radio?</p> <p>Guglielmo Marconi is credited with inventing the first practical radio system.</p>
<p>point cloud input original mesh (only for reference)</p>  <p>What is its working principle?</p> <p>The fan operates on the principle of converting electrical energy into kinetic energy, which is then used to circulate air.</p> <p>What factors will affect its noise?</p> <p>The fan's noise is influenced by its speed, the number of blades, and the material it's made of.</p> <p>Do you have any suggestions on how to reduce noise?</p> <p>Yes, you can reduce the fan's noise by adjusting its speed or by replacing it with a quieter model.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>What color is it? How many wings does it have?</p> <p>The butterfly is white and it has four wings.</p> <p>What role does it play in the ecosystem?</p> <p>The butterfly plays a crucial role in pollination, aiding in the reproduction of flowering plants.</p> <p>What is its relationship with caterpillars?</p> <p>The caterpillar is the larval stage of the butterfly, and it undergoes metamorphosis to become a butterfly.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>How many people can sit here?</p> <p>The bicycle can accommodate two people.</p> <p>What is its working principle?</p> <p>The bicycle is powered by pedals and a chain, which are connected to the wheels.</p> <p>What is the difference between an electric bicycle and it?</p> <p>An electric bicycle is a bicycle that is powered by an electric motor instead of pedals.</p>	<p>point cloud input original mesh (only for reference)</p>  <p>Why is it round?</p> <p>The round shape of the tire is a common design feature in most vehicles. It allows for a smooth ride and better grip on the road.</p> <p>What is the help of the friction force experienced by the tire during movement?</p> <p>The friction force experienced by the tire during movement is what allows the tire to grip the road and provide traction.</p> <p>Can the car not have tires?</p> <p>No, a car cannot function without tires. Tires are essential for the car to move and provide a smooth ride.</p>

Figure 3: 3D question answering results of our MiniGPT-3D. MiniGPT-3D supports multi-round conversation regarding the 3D object. The correct and fine-grained answers are highlight in green.

Table 1: Qualitative results on ModelNet40. The first image in each sample serves as the input for image-based models, while additional views are provided for reference purposes. Compared to other methods, MiniGPT-3D demonstrates superior accuracy in identifying object categories, capturing richer information, and generating more comprehensive responses.




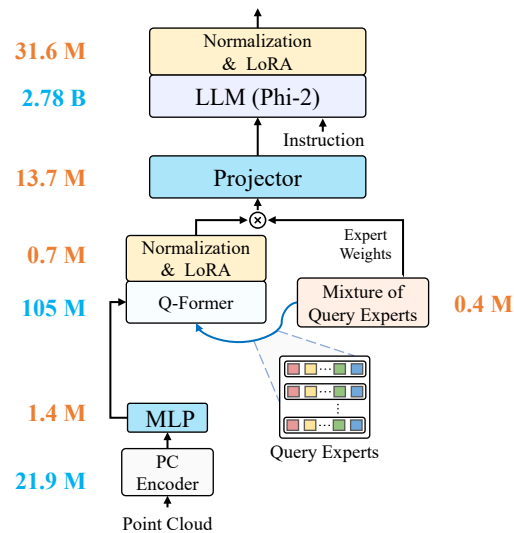
Samples 1, 2		
Prompt	What is this?	This is an object of
Ground Truth	Sofa	Bed
InstructBLIP-13B [1]	sofa	person cutting meat into small pieces with a large knife
LLaVA-13B[5]	The image is a black and white drawing of a couch.	The image is a black and white depiction of a chair.
Point-Bind LLM [2]	This is a drawing of a person laying on a couch, with a dog nearby.	This is an object of art.
PointLLM-13B [7]	This 3D object is a comfortable, black, three-seater sofa.	This is a model of a neatly made bed, primarily composed of a solid black mattress.
MiniGPT-3D	This is a 3D model of a modern, sleek black sofa . The sofa is designed with a minimalist aesthetic, featuring clean lines and a neutral color palette. It is likely to be used in contemporary interior design, offering comfort and style .	A 3D model of a bed with a black mattress . The bed is characterized by its rectangular shape and the mattress, which is black in color . The bed has a headboard and footboard , both of which are also black. This model can be used in various digital platforms such as video games, animations, or virtual reality experiences .
Samples 3, 4		
Prompt	What is this?	This is an object of
Ground Truth	Guitar	Bottle
InstructBLIP-13B [1]	telescope	wine bottle
LLaVA-13B[5]	The image is a close-up view of a black ink pen sitting on a white surface.	This image features a black and white depiction of a bottle.
Point-Bind LLM [2]	This is a bird flying in the sky.	This is an object of a person holding a bottle of wine.
PointLLM-13B [7]	This 3D model represents a black electric guitar equipped with a distinctive headstock.	This is a 3D model of a bottle, possibly intended for liquids.
MiniGPT-3D	This is a 3d model of a cartoon-style guitar . the guitar is predominantly black in color , with a distinctively large headstock and a long neck . the body of the guitar is also black, and it has a unique design with a large, flat headstock and a long neck. the guitar is likely to be used in animated films, video games, or other digital media.	A 3D model of a bottle , predominantly black in color . The bottle is cylindrical in shape, with a narrow neck and a wide base . The bottle appears to be made of plastic or glass , which is a common material used in the production of bottles. The bottle could be used for various purposes such as storing liquids, serving beverages, or as a decorative item .
Samples 5, 6		
Prompt	What is this?	This is an object of
Ground Truth	Bathtub	Monitor
InstructBLIP-13B [1]	table	the foot
LLaVA-13B[5]	The image features a black and white, close-up view of a grater, which is an object used for grating cheese or other food items.	The image features a small table with a black base and a white or light-colored top.
Point-Bind LLM [2]	This is a bowl or a vase, which is oval-shaped and has a smooth surface.	This is an object of art, a sculpture, or a decorative piece.
PointLLM-13B [7]	The 3D object model represents a realistic, dark grey bathtub.	This is a 3D model of a desktop computer monitor.
MiniGPT-3D	This is a 3D model of a cartoon-style bathtub . The bathtub is predominantly black in color and has a unique design with a curved shape . It is likely to be used in animated films, video games, or children's content.	A 3D model of a sleek, modern black computer monitor . The monitor is rectangular in shape and has a glossy finish , giving it a sleek and professional look. It is equipped with a high-resolution display, suggesting it's suitable for tasks that require high-quality visuals, such as graphic design, video editing, or gaming . The monitor is mounted on a sturdy stand, providing stability and convenience .

Table 2: Detailed training settings.

Setting		Stage I	Stage II	Stage III	Stage IV
Dataset		Point-text Instruction Dataset [7]			
Dataset Types		Brief Caption	Brief Caption	Detailed Caption & Conversation	Detailed Caption & Conversation
Dataset Scale		660 k	660 k	70 k	70 k
Brief Caption	Batch Size	9	9	-	-
	Sample Ratio	1	1	-	-
Detailed Caption	Batch Size	-	-	6	6
	Sample Ratio	-	-	2	2
Single-round Conversation	Batch Size	-	-	10	10
	Sample Ratio	-	-	3	3
Multi-round Conversation	Batch Size	-	-	4	4
	Sample Ratio	-	-	3	3
Max Epoch		1	1	3	1
Iterations Per Training Epoch		70000	70000	10000	10000
Learn Rate Scheduler		linear_warmup_cosine_lr			
Initialized Learn Rate		0.00003	0.00003	0.00001	0.000005
Min Learn Rate		0.00001	0.00001	0.000001	0.000001
Warmup Learn Rate		0.000001	0.000001	0.000001	0.000001
Warmup Steps		7000	7000	3000	1000
Weight decay		0.05	0.05	0.05	0.05
Point Cloud Encoder	Point Number	8192	8192	8192	8192
	Point Group Size	32	32	32	32
	Point Patch	512	512	512	512
	Hidden Size	384	384	384	384
	Head of Attention	6	6	6	6
	Number of Layer	12	12	12	12
Point Cloud Projection Layer	Number of Layer	2	2	2	2
	Dimension	384->768; 768->1408	384->768; 768->1408	384->768; 768->1408	384->768; 768->1408
Mixture of Query Experts	Router Type	-	-	-	Sparse Router [6]
	Top Experts	-	-	-	2
	Number of Query Experts	-	-	-	8
	Number of Expert Router Layer	-	-	-	2
	Dimension of Expert Router Layer	-	-	-	768->256; 256->8
Q-Former	Rank of LoRA	-	8	8	8
	Alpha of LoRA	-	16	16	16
	Number of Layer	12	12	12	12
	Head of Attention	12	12	12	12
	Hidden Size	768	768	768	768
Modality Projector	Number of Layer	2	2	2	2
	Dimension	768->4096; 4096->2560	768->4096; 4096->2560	768->4096; 4096->2560	768->4096; 4096->2560
Large Lanuguage Model Backbone	Rank of LoRA	64	64	64	64
	Alpha of LoRA	16	16	16	16
	Number of Layer	32	32	32	32
	Head of Attention	32	32	32	32
	Hidden Size	2560	2560	2560	2560



(a) Architecture, module parameters of MiniGPT-3D.

Trainable Module	Params	Frozen Module	Params
Point Cloud Projection Layer (MLP)	1.4 M	PC Encoder	21.9 M
Norm & LoRA of Q-Former	0.7 M	Q-Former	105 M
Modality Projector	13.7 M	LLM (Phi-2)	2780 M
Mixture of Query Experts	0.4 M	-	-
Norm & LoRA of LLM	31.6 M	-	-
Total Parameters	47.8 M	-	2907 M

(b) Parameters and trainability of modules in MiniGPT-3D.

Figure 4: Architecture, module parameters, and module trainability of MiniGPT-3D. Blue and orange fonts indicate non-trainable and trainable parameters, respectively.

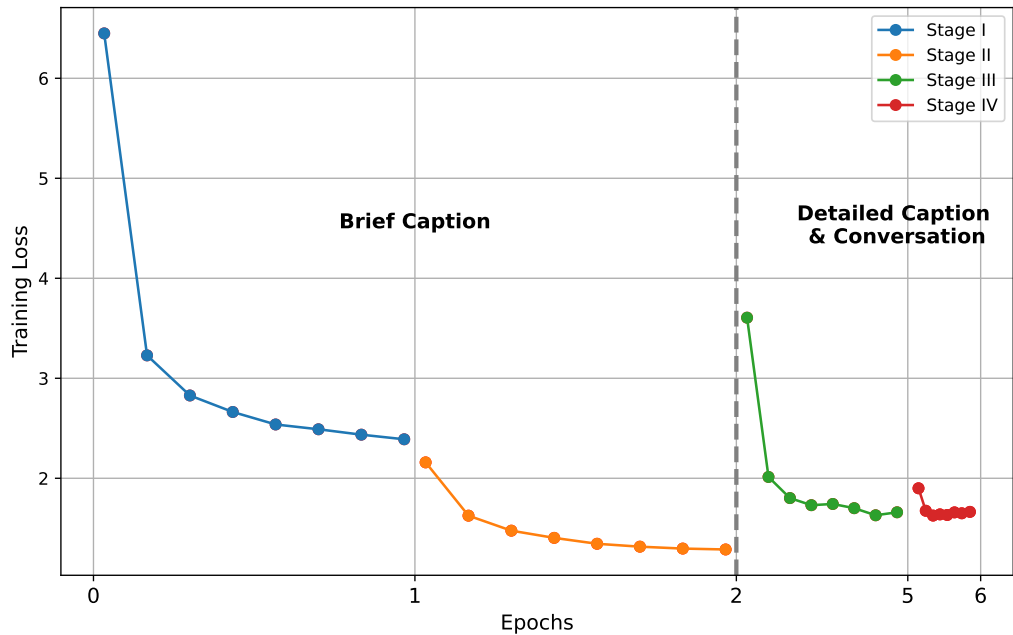


Figure 5: Changes in loss across the four training stages of MiniGPT-3D.

REFERENCES

[1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).

[2] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615* (2023).

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

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language

models. In *International conference on machine learning*. PMLR, 19730–19742.

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[6] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).

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

[8] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19313–19322.

871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928