

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.

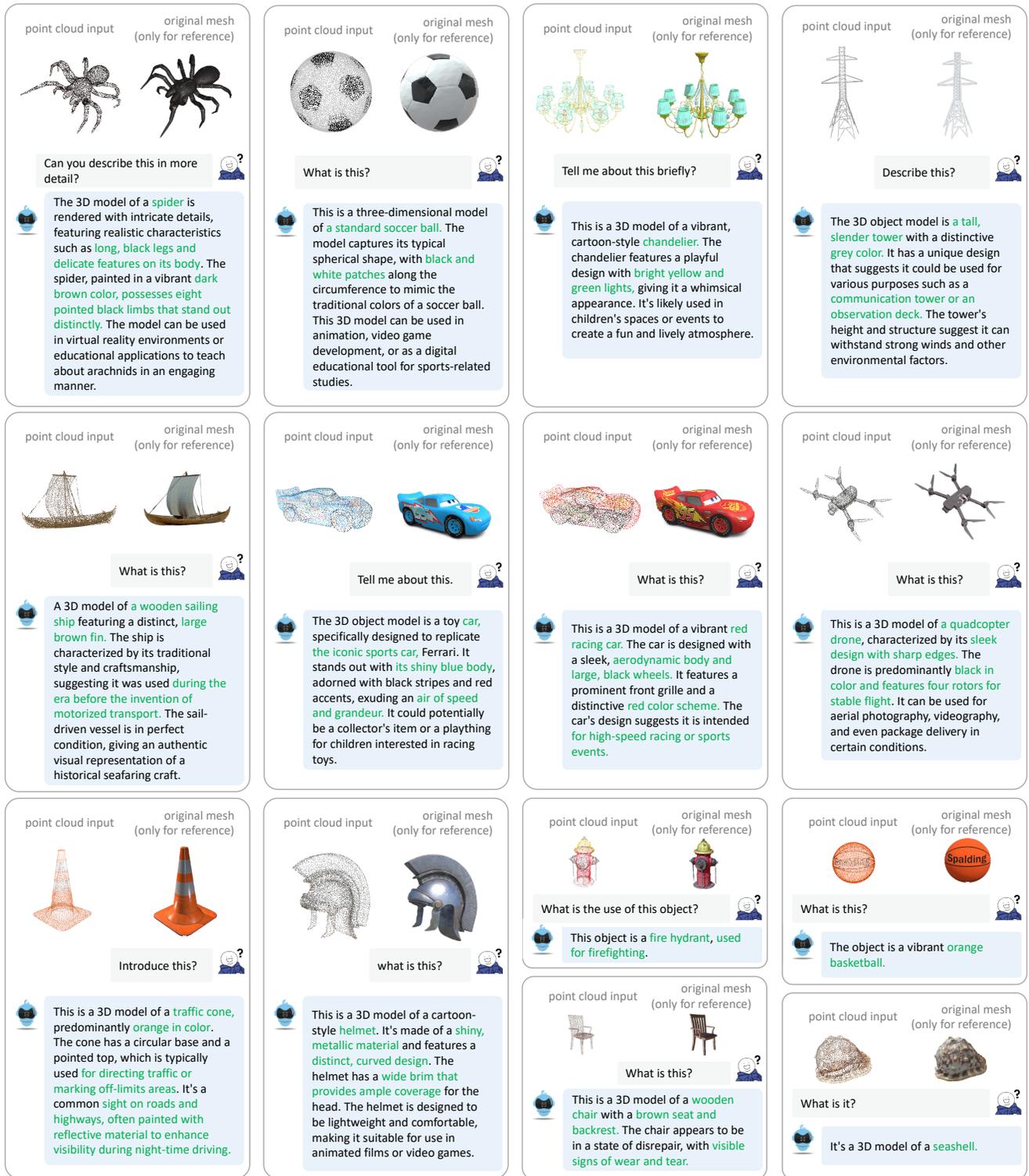


Figure 1: 3D recognition and captioning results (#1) of our MiniGPT-3D. Given a 3D point cloud and a corresponding question, MiniGPT-3D generates a text response. The correct and fine-grained answers are highlight in green.

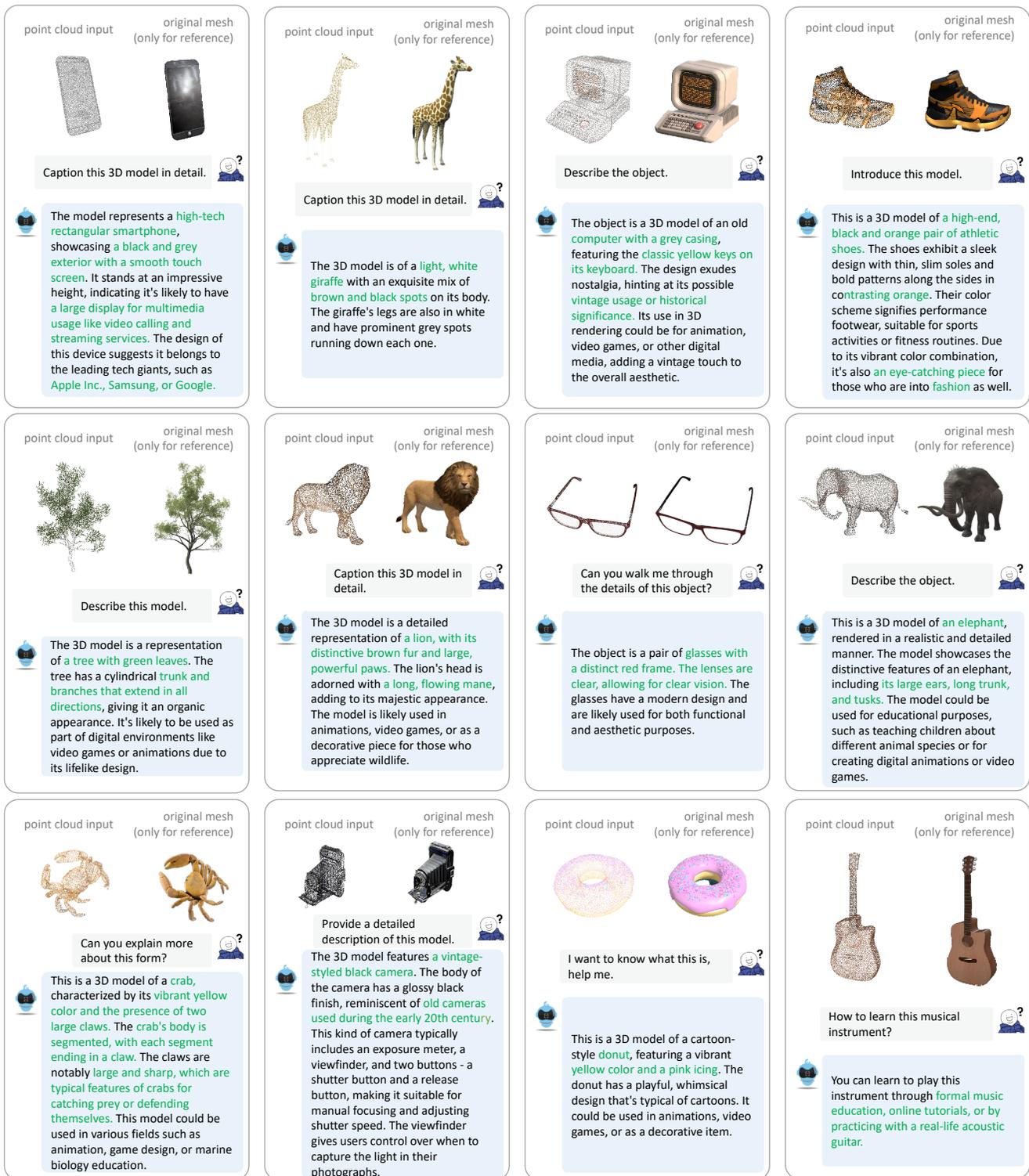


Figure 2: 3D recognition and captioning results (#2) of our MiniGPT-3D. Given a 3D point cloud and a corresponding question, MiniGPT-3D generates a text response. The correct and fine-grained answers are highlight in green.

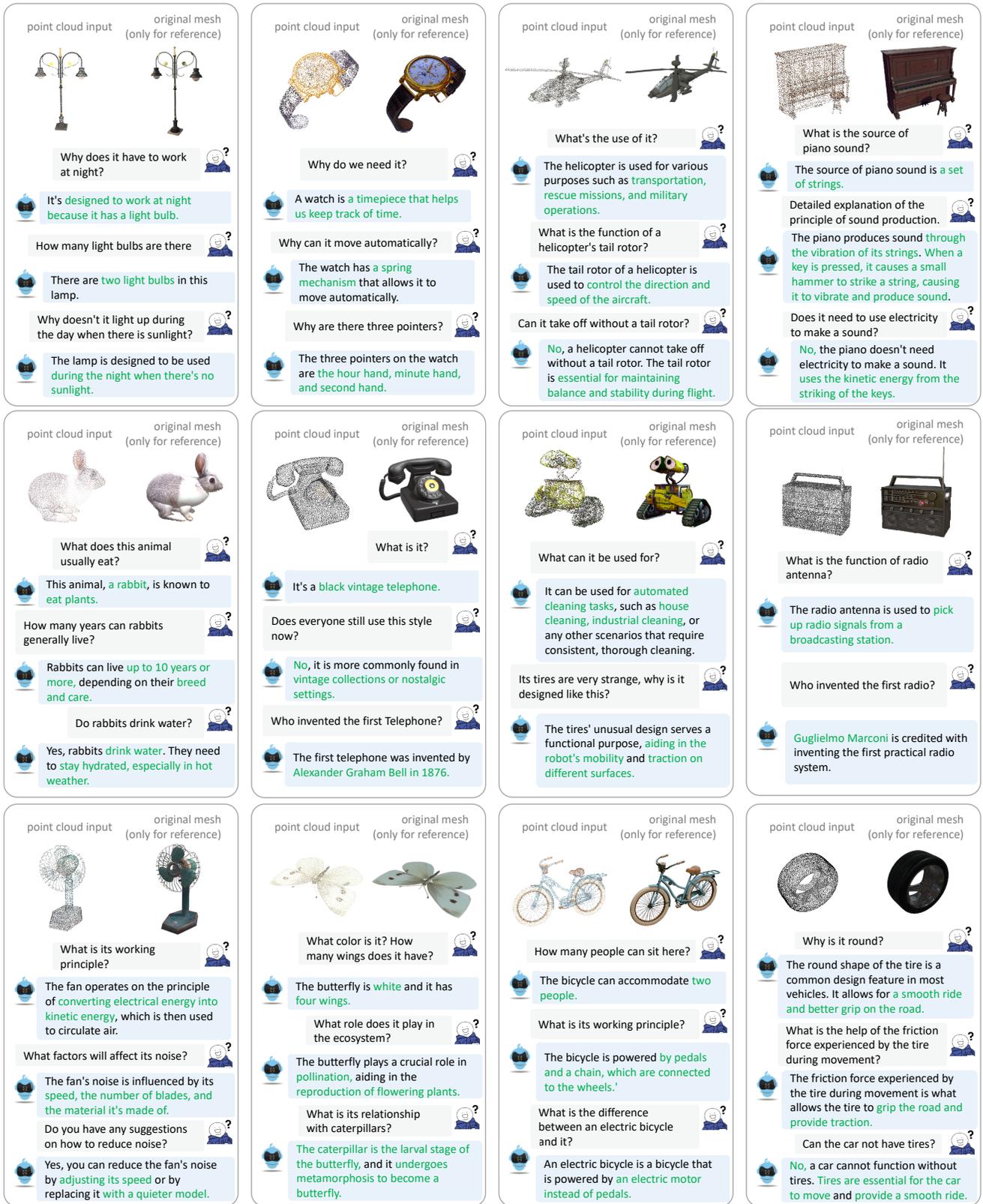
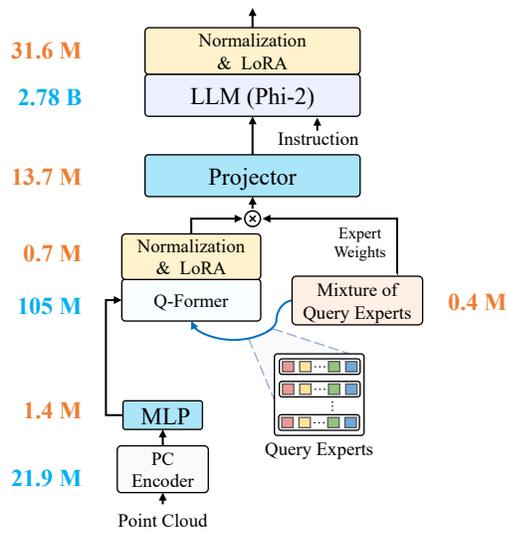


Figure 3: 3D question answering results of our MiniGPT-3D. MiniGPT-3D supports multi-round conversation regarding the 3D object. The correct and fine-grained answers are highlight in green.

Table 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.

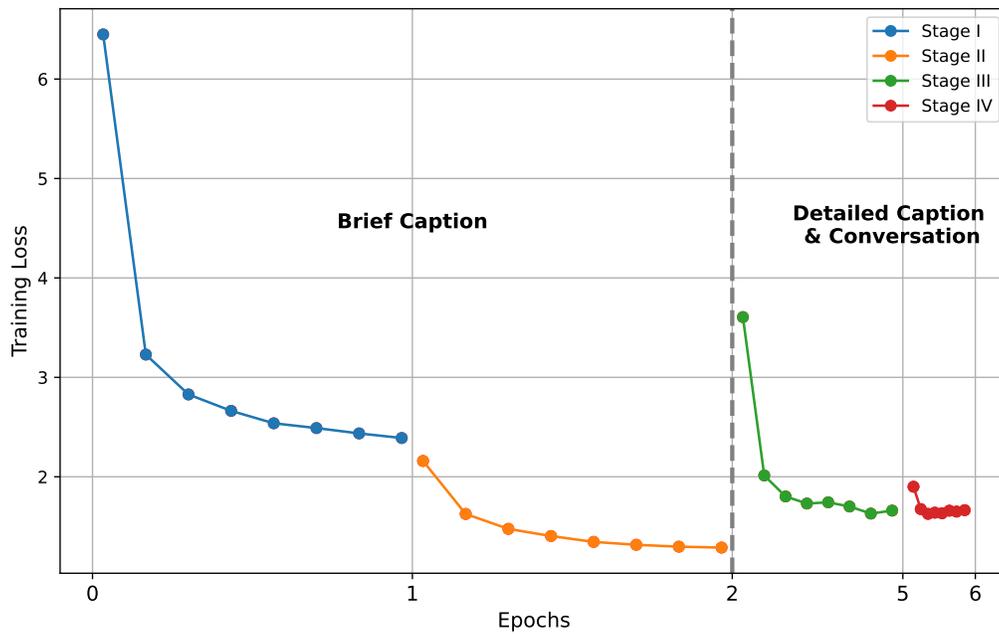


Figure 5: Changes in loss across the four training stages of MiniGPT-3D.

REFERENCES

- 813
- 814 [1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao,
815 Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: To-
816 wards general-purpose vision-language models with instruction tuning. *Advances*
817 *in Neural Information Processing Systems* 36 (2024).
- 818 [2] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming
819 Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. 2023. Point-bind
820 & point-llm: Aligning point cloud with multi-modality for 3d understanding,
821 generation, and instruction following. *arXiv preprint arXiv:2309.00615* (2023).
- 822 [3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean
823 Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large
824 language models. *arXiv preprint arXiv:2106.09685* (2021).
- 825 [4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping
826 language-image pre-training with frozen image encoders and large language
827 models. In *International conference on machine learning*. PMLR, 19730–19742. 871
- 828 [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruc-
829 tion tuning. *Advances in neural information processing systems* 36 (2024). 872
- 830 [6] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Ge-
831 offrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The
832 sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017). 873
- 833 [7] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua
834 Lin. 2023. Pointllm: Empowering large language models to understand point
835 clouds. *arXiv preprint arXiv:2308.16911* (2023). 874
- 836 [8] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu.
837 2022. Point-bert: Pre-training 3d point cloud transformers with masked point
838 modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
839 *recognition*. 19313–19322. 875
- 840 876
- 841 877
- 842 878
- 843 879
- 844 880
- 845 881
- 846 882
- 847 883
- 848 884
- 849 885
- 850 886
- 851 887
- 852 888
- 853 889
- 854 890
- 855 891
- 856 892
- 857 893
- 858 894
- 859 895
- 860 896
- 861 897
- 862 898
- 863 899
- 864 900
- 865 901
- 866 902
- 867 903
- 868 904
- 869 905
- 870 906
- 907
- 908
- 909
- 910
- 911
- 912
- 913
- 914
- 915
- 916
- 917
- 918
- 919
- 920
- 921
- 922
- 923
- 924
- 925
- 926
- 927
- 928