Appendix of UniCode: Learning a Unified Codebook for Multimodal Large Language Models

In this appendix, we delve deeper into both quantitative and qualitative aspects of our experimental outcomes. Firstly, we elaborate on the implementation details of UniCode in Section 1, followed by further discussion of our approach in Section 2. Subsequently, in Section 3, we conduct comparison experiments using additional benchmarks in Visual Question Answering (VQA). Lastly, Section 4 showcases a range of qualitative results derived from image generation tasks across various configurations.

1 Additional Implementation Details

In this study, we utilize Vicuna-7b [2] as the foundational Large Language Model (LLM) for the development of our UniCode. To refine the LLM with greater efficiency, we adopt the LoRA training technique [4]. However, due to computational resource limitations, this research does not explore the use of larger LLMs or engage in full-parameter fine-tuning, with these aspects earmarked for future investigation. Considering the training data, we ensure its compatibility with the LLaVA1.5 format [9]. For tasks related to multimodal understanding and text-driven image generation, our model training integrates the LCS-558K image-text pair dataset [9] with the Conceptual Captions 3M (CC3M) dataset [10]. For unconditioned and class-conditioned generation tasks, UniCode is trained using a combination of LCS-558K data and corresponding downstream datasets from LSUN (cat, bedroom, and church) and ImageNet. We adhere to a training protocol similar to that of LLaVA-1.5, with each training run comprising approximately 6 hours of pretraining followed by 20 hours of visual instruction tuning, utilizing 8 NVIDIA A100 40G GPUs. All experiments in this paper are carried out on 2×8 NVIDIA A100 40G GPUs. It is important to note that our UniCode does not claim superior efficiency in per-token generation compared to the original LLaMA model. Instead, one of its principal advantages lies in its capacity to represent images with a reduced number of tokens, thereby facilitating the processing of extended memories across multiple frames.

2 Additional Discussion

In this paper, we have demonstrated the promising potential of employing a unified codebook for multimodal large language models. However, there are several facets of this approach that we have not been able to comprehensively tackle. We aim to delve deeper into these aspects in our forthcoming research. Below, we provide a brief overview of some of these areas for future exploration.

Imbalance between Image and Text Data Volumes. This observed discrepancy can primarily be attributed to the constrained access to high-quality image captions. For text generation, our instructional dataset, following Liu *et al.* [8], contains only LCS-558K image-text pairs. In stark contrast, the scope for collecting images to enhance a model in image generation tasks seems almost limitless. For instance, 3 million images from CC3M, 1 million from ImageNet, and another 3 million from LSUN-cat categories. A significant challenge we encounter is the effective amalgamation of this extensive image generation data with the comparatively sparse text generation data. In our present endeavor, we have approached this challenge by simply combining the text generation dataset with the corresponding image dataset for each benchmark within image generation tasks. Therefore, the scale of image generation data can be limited.

Limitation of Visual Tokenization in Detail Capture. Our visualization examples of image reconstruction clearly illustrate that existing VAE-style visual tokenizers struggle to capture fine details, especially in text and faces. This shortfall potentially contributes to the significant performance gap observed between UniCode and LLaVA across related benchmarks.

Gap in Linking Visual Codes to Semantic Meaning. In our study, the exploration of how codes within the unified codebook correlate with semantic content remains untouched. Initiatives such as SPAE [11] have suggested leveraging robust image-text matching models like CLIP for semantic guidance, thereby enhancing the semantic richness of the codes. This aspect is not addressed here. Additionally, the existence of a positive correlation between visual and text generation data has not been established in our research. The notable lack of high-quality image captions relative to the abundance of images presents a significant challenge in effectively merging these two types of data for more efficient instruction tuning during practical training. Addressing these issues will be the focus of our future investigations.

Table 1: Comparison with MLLMs on additional VQA benchmarks. UniCode achieves competitive results against other top MLLMs while requiring less data and fewer parameters for its visual tokenizer. Here, "M2T" and "M2M" refer to the model's capability to generate either text only or multiple modalities. "Vis-P", "PT" and "IT" represent the number of parameters in the visual encoder, the number of samples for multimodal alignment, and instruction tuning, respectively.

Method	Type	LLM	Vis-P	PT	IT	GQA	SEED	LLaVA$^{\text{W}}$	MM-Vet
BLIP-2 [7]	M2T	Vicuna-13B	303M	129M	0	41.0	46.4	38.1	22.4
InstructBLIP [3]	M2T	Vicuna-7B	303M	129M	1.2M	49.2	53.4	60.9	26.2
Qwen-VL [1]	M2T	Qwen-7B	1.8B	1.4B	50M	59.3	56.3	-	-
LLaVA-1.5	M2T	Vicuna-7B	303M	558K	665K	62.0	61.6	70.7	35.4
UniCode	M2M	Vicuna-7B	104M	0	665K	44.6	38.3	52.1	12.7
UniCode+	M2M	Vicuna-7B	1B	0	665K	50.6	48.1	62.7	23.1

3 Additional Experimental Results

We extend our experimental comparison to additional VQA benchmarks, as detailed in Table 1. These benchmarks feature diverse datasets, including GQA [5], SEED-Bench [6] (SEED), LLaVA-Bench-in-the-Wild (LLaVA$^{\mathrm{W}}$) [9], and MM-Vet [12]. The outcomes from these comparisons clearly demonstrate that enhancements in visual configuration significantly boost performance across these benchmarks. Such results certainly highlight the effectiveness and the considerable potential of our UniCode in the realm of multimodal understanding.

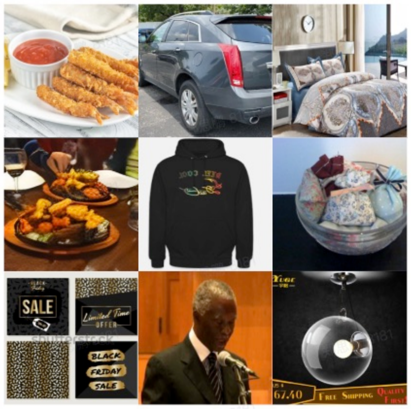


Fig. 1: Raw images of image reconstruction examples in our main paper.

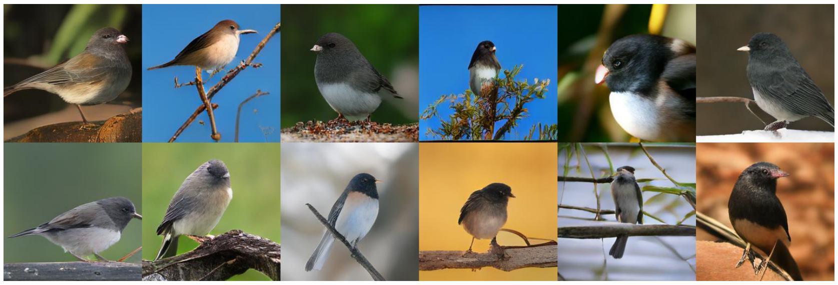


Fig. 2: Qualitative examples of class-conditioned image generation on the category "junco" in ImageNet.

4 Additional Qualitative Examples

In our main paper, we showcase qualitative examples to highlight the image reconstruction capabilities of our proposed UniCode. For a comparative analysis, Figure 1 displays the original, unaltered images. It is evident that current VAE tokenizers falter in accurately capturing intricate details, notably in text and facial features. Furthermore, we present a selection of visualizations demonstrating the outcomes of **class-conditioned image generation** on ImageNet, specifically featured in Figures 2, 3, 4, and 5 for a variety of categories. In addition, we showcase **unconditioned image generation** results on LSUN-{cat, bedroom, church} as depicted in Figures 6, 7, and 8, respectively. Lastly, we offer illustrations of **text-conditioned image generation** using the CC3M dataset in Figure 9, in order to further exemplify the versatility of UniCode's generative capabilities.

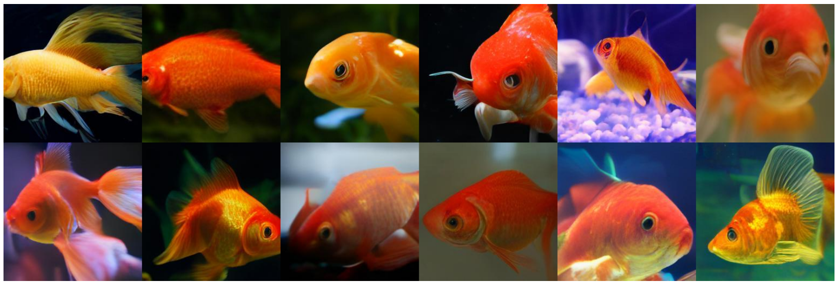


Fig. 3: Qualitative examples of class-conditioned image generation task on the category "goldfish" in ImageNet.

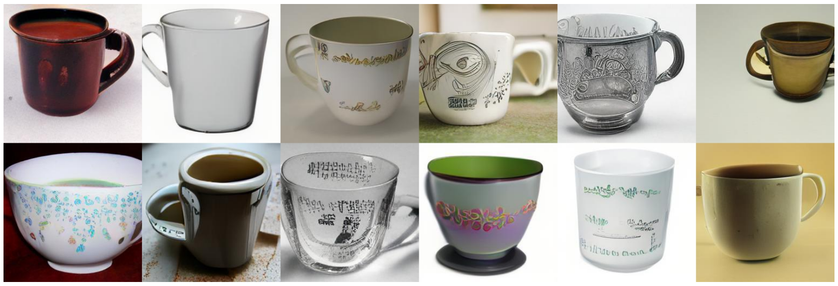


Fig. 4: Qualitative examples of class-conditioned image generation task on the category "cup" in ImageNet.

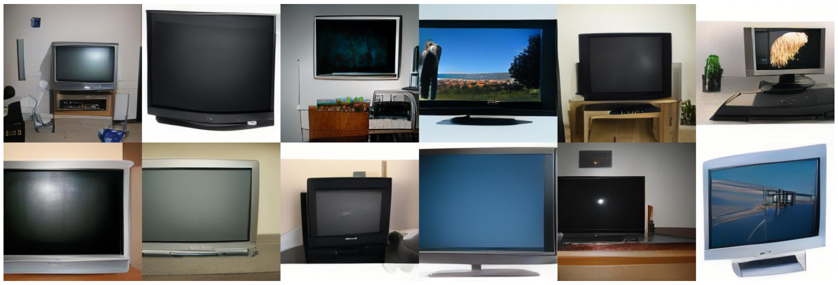


Fig. 5: Qualitative examples of class-conditioned image generation task on the category "television" in ImageNet.

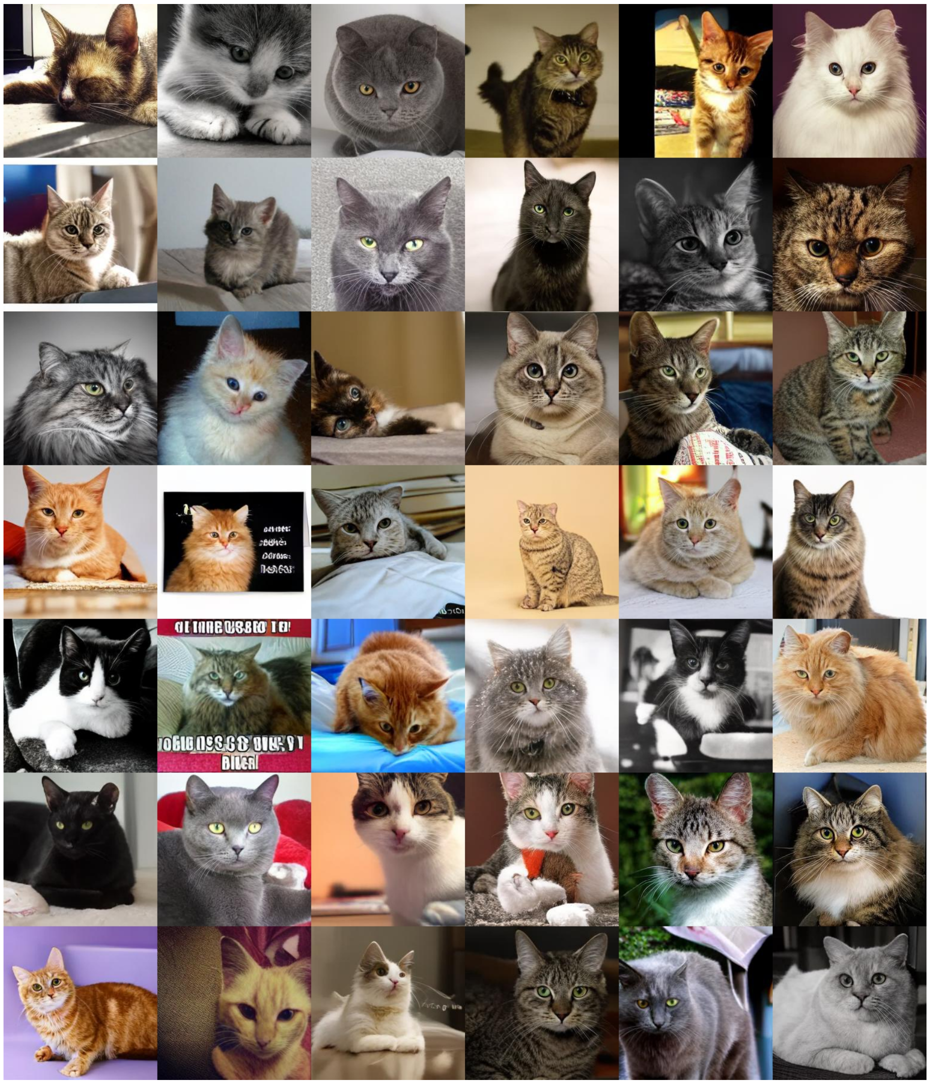


Fig. 6: Qualitative examples of unconditional image generation task on LSUN-Cat.

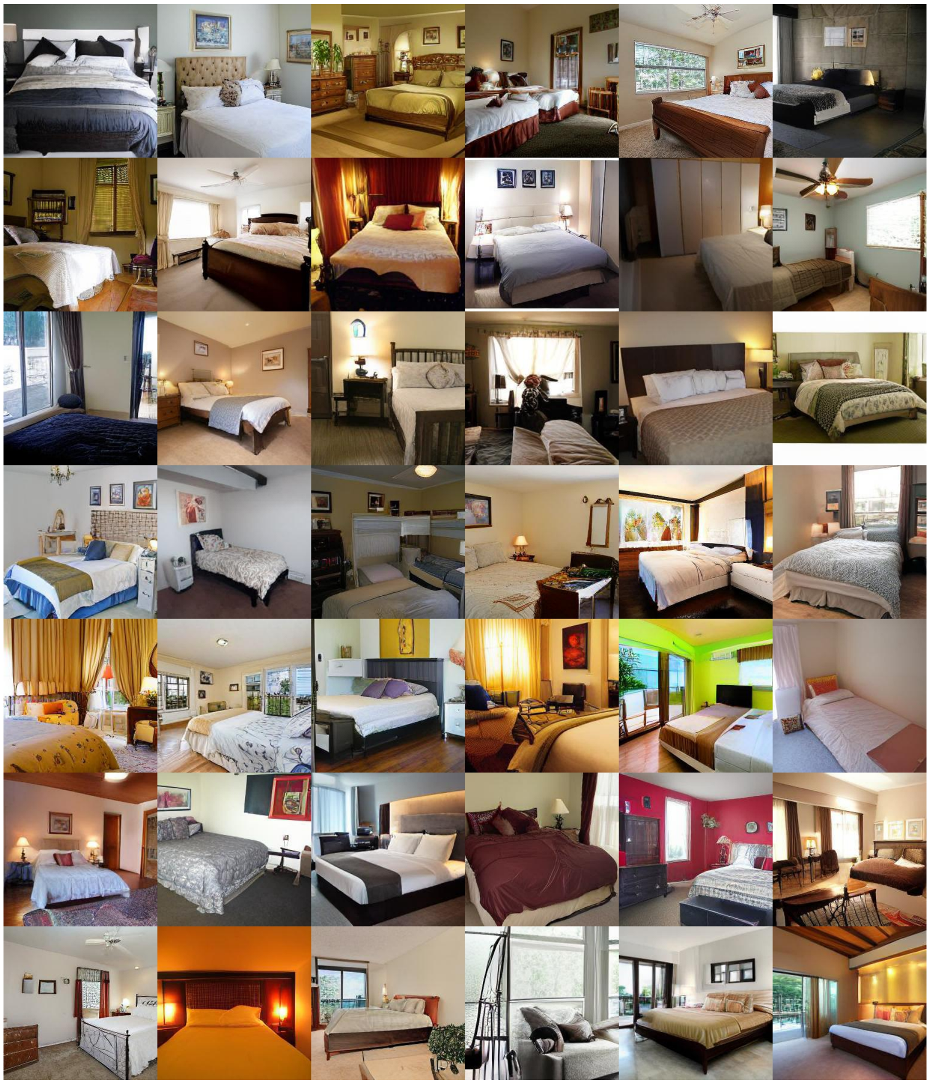


Fig. 7: Qualitative examples of unconditional image generation task on LSUN-Bedroom.

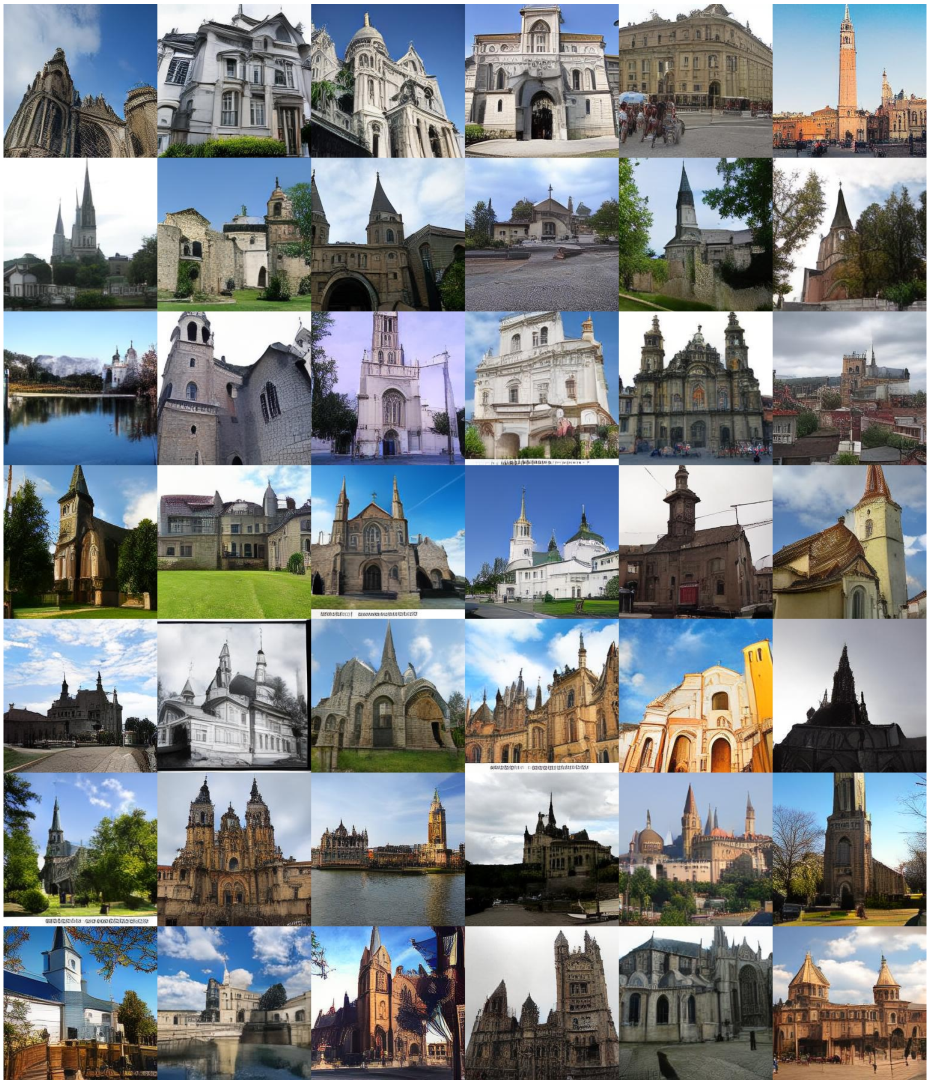
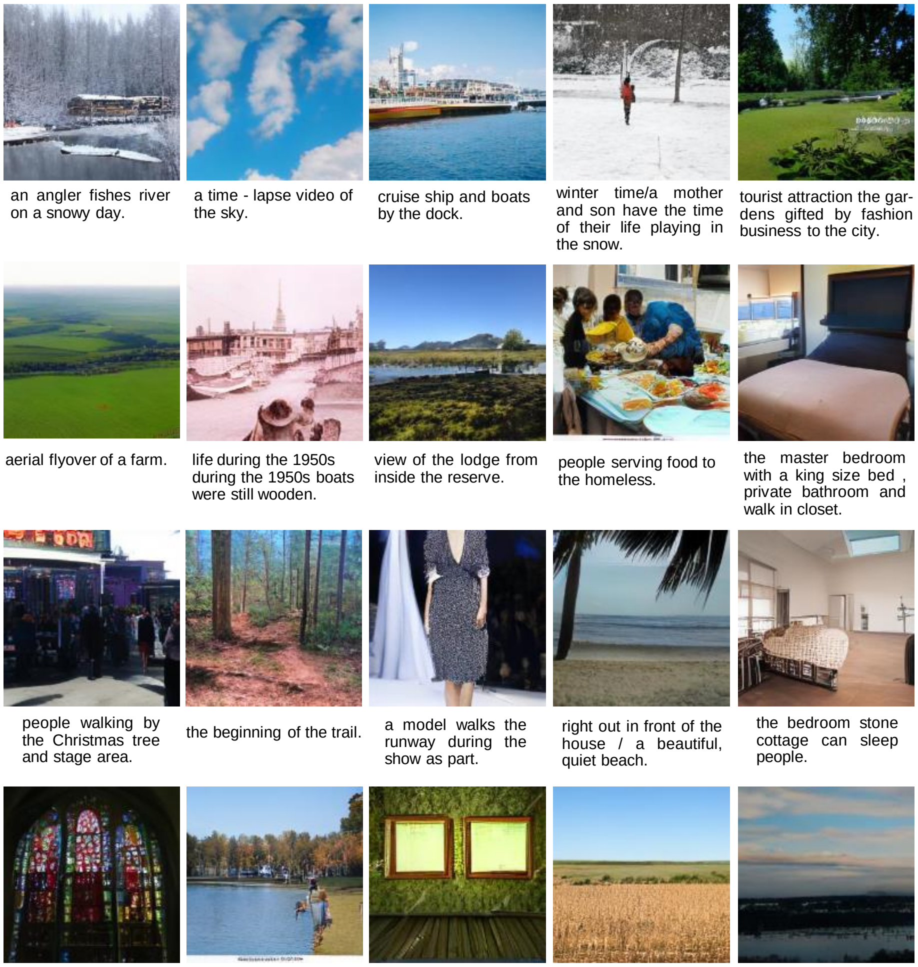


Fig. 8: Qualitative examples of unconditional image generation task on LSUN-Church.



an angler fishes river on a snowy day.

a time - lapse video of the sky.

cruise ship and boats by the dock.

winter time/a mother and son have the time of their life playing in the snow.

tourist attraction the gardens gifted by fashion business to the city.

aerial flyover of a farm.

life during the 1950s during the 1950s boats were still wooden.

view of the lodge from inside the reserve.

people serving food to the homeless.

the master bedroom with a king size bed , private bathroom and walk in closet.

people walking by the Christmas tree and stage area.

the beginning of the trail.

a model walks the runway during the show as part.

right out in front of the house / a beautiful, quiet beach.

the bedroom stone cottage can sleep people.

basilica of roman cath olic place of worship / stained glass window.

people on the shore of lake.

gold frames on the wall with green wallpaper.

dry fields on a summer day.

clouds moving rapidly in blue skies reflecting sunrise, then vanishing to clear skies.

Fig. 9: Qualitative examples of text-conditioned image generation task on CC3M.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
2. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
3. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
4. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
5. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
6. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
7. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
8. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
9. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
10. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
11. Yu, L., Cheng, Y., Wang, Z., Kumar, V., Macherey, W., Huang, Y., Ross, D.A., Essa, I., Bisk, Y., Yang, M.H., et al.: Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. arXiv preprint arXiv:2306.17842 (2023)
12. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)