LangBridge: Interpreting Image as a Combination of Language Embeddings

Supplementary Material

Overview

In this supplementary material, we present more dataset details and more experimental results that are not included in the main paper. The contents include:

- A comprehensive introduction to evaluation benchmarks S-A.
- Detailed training configurations and hyperparameters S-B.
- Additional experimental results on a broader range of benchmarks for LLaVA-Next integrated with LangBridge. S-C
- Discussion of Computational Cost. S-D
- Discussion of current limitations and future directions S-E.

S-A. Evaluation Benchmarks

We evaluated our method on seven widely-used benchmarks, covering a diverse range of vision-language understanding tasks.

- **GQA [1]**: Evaluates the model's visual perception ability through open-ended questions.
- **TextVQA [6]**: Tests the model's ability to read and reason about text in images to answer questions, focusing on text-based visual reasoning.
- **ScienceQA [5]**: Provides a set of multiple-choice science questions with images to test the model's zero-shot generalization ability in scientific question answering.
- **MME [7]**: A comprehensive evaluation of LVLMs across ten perception tasks (e.g., OCR and object recognition) and four cognitive tasks (e.g., commonsense reasoning, numerical computation, translation, and code reasoning).
- **MMBench [4]**: A bilingual benchmark for evaluating LVLM's multimodal understanding capabilities, consisting of approximately 3000 multiple-choice questions covering 20 ability dimensions. The Chinese version is called MMBench-CN.
- **MMVeT [8]**: A challenging multimodal benchmark designed to evaluate vision-language models' robustness and reliability. It focuses on testing fine-grained visual understanding, complex reasoning, and real-world application scenarios.
- **POPE [2]**: Evaluates the model's ability to identify specific objects in images, aiming to detect object-level hallucinations. It uses "yes/no" questions based on object annotations, with 50% of queries targeting existing objects and 50% targeting non-existing objects, employing random, popular, and adversarial sampling strategies.

Through these comprehensive benchmarks, we systematically evaluated the model's capabilities across diverse tasks, with particular emphasis on multimodal understanding, visual reasoning, hallucination detection, and real-world applicability. These benchmarks collectively provide a thorough assessment of the model's strengths and potential areas for improvement.

S-B. Training Details

As shown in Table S-1, we follow the same training recipes as LLaVA's settings [3] for standard MLPs, except that we change the pretrain learning rate from 1e-3 to 2e-5 for LangBridge.

Table S-1. Training hyper-parameters

Hyper-parameter	Value
batch size	256 (pretrain), 128 (finetune)
learning rate	1e-3 (pretrain), 2e-5 (finetune)
learning rate schedule	cosine
learning rate warm-up ratio	0.03
weight decay	0
epoch	1
optimizer	AdamW
float precision	bfloat16
deepspeed configuration	zero2 (pretrain), zero3 (finetune)

S-C. More experiments results

To further evaluate the generalizability and robustness of our method, we integrate it into LLaVA-Next. Our model first uses a Qwen2-0.5B pretrained LangBridge module, which is then combined with Qwen2-7B for further supervised fine-tuning (SFT). We compare it to a standard MLP baseline with same backbone and evaluate across a range of benchmarks. As shown in Table S-2, our model achieves consistent improvements over the baseline in fine-grained image analysis tasks, including AI2D, ChartQA, and DocVQA, while maintaining parity on MMVP. For grounding and real-world reasoning tasks (Table S-3), our approach significantly outperforms the baseline on RefCOCO (+17.06%) and RefCOCO+ (+15.23%), and performs comparably on RefCOCOg and MMRealWorld. In the video understanding domain (Table S-4), the model improves results on Seed-Video and Seed2-Video while maintaining similar performance on VideoMME and MMT. Lastly, as shown in Table S-5, our method improves accu-

racy on TextVQA, GQA, and MME, demonstrating its effectiveness on widely used benchmarks, with only a slight drop on MMMU. These results collectively demonstrate that LangBridge consistently matches or outperforms standard MLPs across a wide range of tasks, showcasing strong generalization capability.

Table S-2. Results on Fine-grained Benchmarks

Method	AI2D	ChartQA	DocVQA	MMVP
Baseline	0.763	0.8632	0.741	43.3
Our Model	0.766 (100.39%)	0.877 (101.60%)	0.758 (102.29%)	43.3 (100.00%)

Table S-3. Results on Grounding and Real-World Benchmarks

Method	RefCOCO	RefCOCO+	RefCOCOg	MMRealWorld
Baseline	0.387	0.348	0.6377	0.4405
Our Model	0.453 (117.06%)	0.401 (115.23%)	0.622 (97.53%)	0.4436 (100.70%)

Table S-4. Results on Video Benchmarks

Method	VideoMME	Seed-Video	Seed2-Video	MMT
Baseline	50.77	0.444	0.494	54.43
Our Model	50.71 (99.88%)	0.450 (101.35%)	0.504 (102.03%)	53.98 (99.17%)

Table S-5. Results on Commonly Used Datasets

Method	TextVQA	GQA	MME	MMMU
Baseline	0.654	0.651	1886	0.4356
Our Model	0.665 (101.68%)	0.652 (100.15%)	1928 (102.23%)	0.4178 (95.91%)

S-D. Computational Cost

We follow LLaVA-Next settings and conduct experiments on H100 using its SFT data. As shown in the table below, LangBridge incurs only a 10% increase in training time compared to the baseline, while significantly reducing pretraining time.

Vision	Backbone LLM	MLP	Speed (s/iter)	Time (h)	Parameters	Data	Hardware
CLIP-L/336	Qwen2-7B-Instruct	Normal	3.876	24.8	7.3B	LLaVA1.6 SFT	8× H100
CLIP-L/336	Qwen2-7B-Instruct	Langbridge	4.273 (1.1×)	27.3	7.43B	LLaVA1.6 SFT	8× H100

S-E. Limitations

While our work demonstrates promising results in vision-language tasks, it has a notable limitation: we only focus on the visual modality. Modern multimodal systems often need to process various types of inputs beyond images, such as videos, audio, and 3D data. Future work could explore extending LangBridge to support these additional modalities, potentially enabling a more comprehensive multimodal understanding system.

References

[1] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 1

[2] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[4] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[5] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 1

[6] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1

[7] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[8] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1