Less is More: A Simple yet Effective Token Reduction Method for Efficient Multi-modal LLMs

Dingjie Song[♥,♠], **Wenjun Wang**[♥], **Shunian Chen**[♥], **Xidong Wang**[♥], **Michael Guan**[♥], **Benyou Wang**[♥*]
[♥]The Chinese University of Hong Kong, Shenzhen [♠]Lehigh University
https://github.com/FreedomIntelligence/TRIM/

Abstract

The rapid advancement of Multimodal Large Language Models (MLLMs) has led to remarkable performances across various domains. However, this progress is accompanied by a substantial surge in the resource consumption of these models. We address this pressing issue by introducing a new approach, Token Reduction using CLIP Metric (TRIM), aimed at improving the efficiency of MLLMs without sacrificing their performance. Inspired by human attention patterns in Visual Question Answering (VQA) tasks, TRIM presents a fresh perspective on the selection and reduction of image tokens. The TRIM method has been extensively tested across 12 datasets, and the results demonstrate a significant reduction in computational overhead while maintaining a consistent level of performance. This research marks a critical stride in efficient MLLM development, promoting greater accessibility and sustainability of high-performing models.

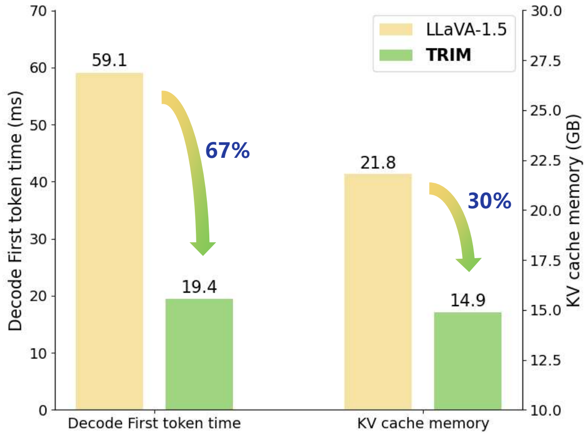

Figure 1: Comparison of First Token Decoding Time and KV Cache Memory Usage (FP16, batch size=1) for LLaVA-1.5 between the baseline and TRIM, where each image is accompanied by an average 40-token question.

1 Introduction

The rapid development of MLLMs has demonstrated superior, and sometimes even superhuman, performance across various fields (OpenAI, 2023; Team et al., 2023; Liu et al., 2023a, 2024b; Song et al., 2024a; Ge et al., 2024; Song et al., 2024b). However, this progress comes with a significant increase in the resources consumed by these models. As a result, the research community has begun to place a greater emphasis on developing efficient MLLMs (Jin et al., 2024; Xu et al., 2024a).

Current efforts include developing lighter architectures to reduce parameters and computational complexity (Lin et al., 2024; Zhao et al., 2024; Chen et al., 2024; Yuan et al., 2023), creating specialized components to optimize efficiency and add properties like locality (Chu et al., 2024; Cha et al., 2024), and, notably, enhancing support for resource-intensive tasks through techniques such as visual token compression. Visual token compression reduces the number of tokens needed to represent visual data, thereby lowering computational and memory demands without sacrificing performance. This approach is particularly crucial as it enables the efficient processing of high-resolution images and videos (Xu et al., 2024b; Gao et al., 2024; Wang et al., 2024b).

Before the MLLM era, numerous efforts aimed to reduce the number of tokens. For instance, methods like MADTP (Cao et al., 2024) were proposed, but they did not integrate closely enough with Large Language Models (LLMs). In the context of MLLMs, the only notable work is PruMerge (Shang et al., 2024), which uses self-attention in vision encoder to make judgments; however, it remains a sub-optimal method for deciding which tokens to reduce.

Drawing inspiration from human attention patterns (Prat-Ortega and de la Rocha, 2018) in VQA tasks, our proposed method employs the use of CLIP (Radford et al., 2021) representations to

*Corresponding author. Email: wangbenyou@cuhk.edu.cn

Proceedings of the 31st International Conference on Computational Linguistics, pages 7614–7623
January 19–24, 2025.

calculate the similarity between text and image patches. Through our observations, we found that this similarity metric effectively identifies semantically relevant regions within images.

Building on this foundation, we introduce an innovative approach known as **TRIM** (**T**oken **R**eduction using CL**I**P **M**etric). In this method, the CLIP metric is leveraged to evaluate the significance of each image token. We also propose to use the Interquartile Range (IQR) (Boukerche et al., 2020) scoring function that adaptively selects image tokens integral to question answering. To account for potential information loss, the selected image tokens are supplemented with an aggregated token that preserves information from the non-selected tokens. This methodology significantly streamlines the computational process, reducing the number of image tokens by approximately 79%, processing time by 67%, and memory usage by 30% relative to the baseline, as depicted in Figure 1. Importantly, it achieves this efficiency while preserving performance comparable to that of the original model.

Our contributions can be summarized as follows:

- We observed that the CLIP metric can effectively capture important image tokens.
- By leveraging the CLIP metric and the IQR scoring function, we adaptively select image tokens that are crucial for answering questions, while an aggregated token is used to retain additional image information.
- Extensive testing on 12 datasets demonstrates that our TRIM method significantly reduces computational overhead while maintaining consistent performance.

2 Related Work

Many works focus on better projecting visual information into the text embedding space. Early work (Alayrac et al., 2022) uses a perceiver resampler to integrate visual data into the language model's hidden layers. Some works (Li et al., 2023b; Zhu et al., 2023; Bai et al., 2023; Li et al., 2024; Jian et al., 2024) compress visual tokens to a fixed length and map them to text space using linear layers. More recent methods using the LLaVA architecture (Liu et al., 2024a; AI et al., 2024; Wang et al., 2024a; Zhu et al., 2024; Chen et al., 2024) simplify this by using MLP layers to map visual tokens to text space, reducing training parameters and data requirements, thus gaining popularity for their efficiency and simplicity.

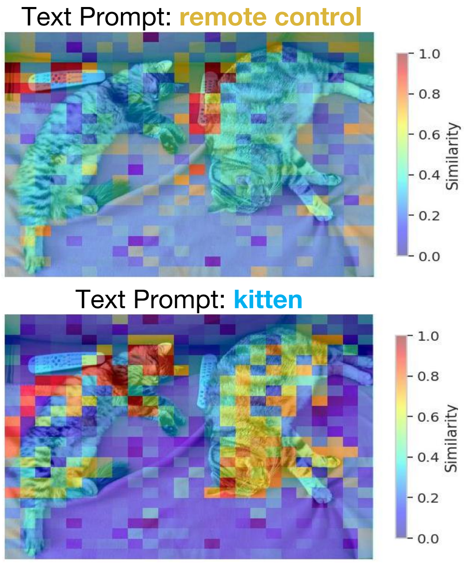

Figure 2: Similarity map between different text prompts and image patch representations within an image. Representations are extracted from the CLIP, and similarity between text and image patches is computed using dot product (CLIP metric). This demonstrates the CLIP metric's ability to capture text-image patch relationships.

However, LLaVA encounters increased computational load in multi-image scenarios due to the high number of visual tokens encoded by standard CLIP (Radford et al., 2021; Vaswani et al., 2023). Compressing these tokens while retaining visual information is crucial. Though traditional CV tasks have used token merging and pruning effectively (Rao et al., 2021; Meng et al., 2021; Cao et al., 2023), this approach is underexplored in MLLMs, where direct adoption of their method is more time-consuming. LLaVA-PruMerge (Shang et al., 2024) recently attempted token reduction using `[CLS]` token-based similarity with sub-optimal results. In contrast, we introduce a token reduction method based on the similarity of text and visual tokens, achieving comparable performance with significantly fewer visual tokens.

3 Method

3.1 Observations

One of the challenges in token reduction is determining the importance of different tokens. As selective attention theory (Prat-Ortega and de la Rocha, 2018) describes, selective attention in human vision prioritizes focal areas, enabling detailed processing while disregarding irrelevant informa-

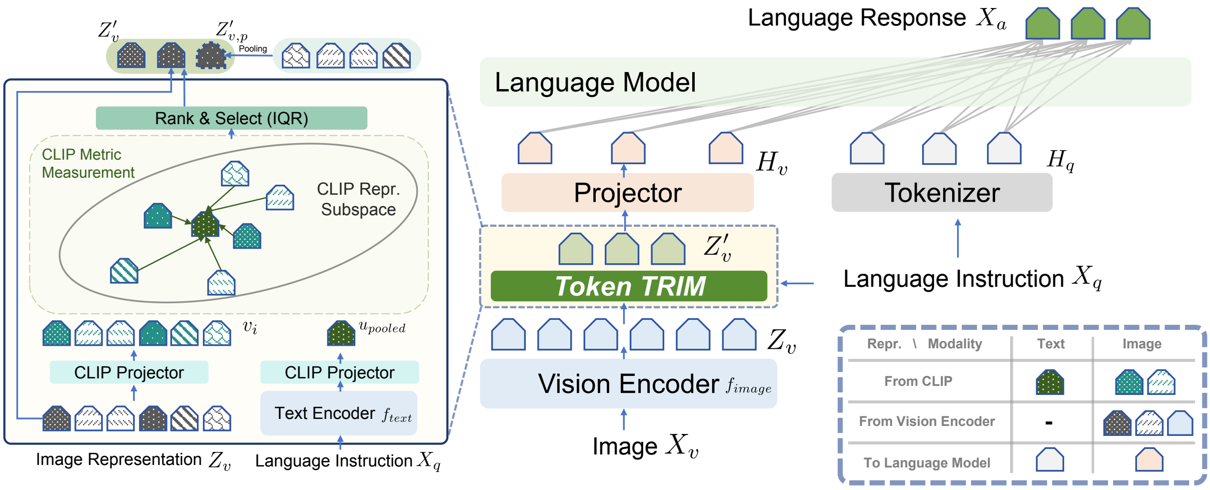

Figure 3: Overview of TRIM and LLaVA architecture. TRIM involves three steps: (1) Calculating the similarities between the text token and visual tokens; (2) Ranking and selecting the important tokens based on these similarities using an outlier detection algorithm; (3) Appending an aggregated token derived from the unselected image tokens.

tion. To simulate this attention mechanism, we need to establish a connection between the text and image patches. We observed that the CLIP model, during its training process, implicitly establishes such connections. CLIP uses contrastive learning loss to bring matching text-image pairs closer and push non-matching pairs apart. By leveraging these representations, we can compute and analyze the similarity between text representations and image patch representations. As depicted in Figure 2, we found that by using text representations, the similarity metric effectively captured semantically relevant image patches.

3.2 Token Reduction with TRIM

Building upon the observations, we put forth a novel token reduction method coined as TRIM (**T**oken **R**eduction using CLIP **M**etric), as depicted in Figure 3, which primarily consists of three steps.

Assessing Token Significance. First, we harness the similarity metric from CLIP to assess the significance of image tokens. Given the feature representations extracted from the text encoder f_{text} and the image encoder f_{image}, we proceed to calculate the cosine similarity between each image token $\mathbf{v}_i$ and the pooled text representation $\mathbf{u}_{\text{pooled}}$, derived from `[eot]` token in CLIP, as follows:

$$S(\mathbf{v}_i, \mathbf{u}_{\text{pooled}}) = \frac{\mathbf{v}_i \cdot \mathbf{u}_{\text{pooled}}}{\|\mathbf{v}_i\| \|\mathbf{u}_{\text{pooled}}\|}$$

Subsequently, we apply a softmax function to the calculated similarities, yielding:

$$S_{\text{softmax}}(\mathbf{v}_i, \mathbf{u}_{\text{pooled}}) = \frac{e^{S(\mathbf{v}_i, \mathbf{u}_{\text{pooled}})}}{\sum_j e^{S(\mathbf{v}_j, \mathbf{u}_{\text{pooled}})}}$$

This softmax score, $S_{\text{softmax}}(\mathbf{v}_i, \mathbf{u}_{\text{pooled}})$, effectively quantifies the significance of each image token $\mathbf{v}_i$, thereby forming the underlying basis for the token reduction in our method.

Selecting Important Tokens. In order to determine the optimal number of image tokens to retain, we adopt the Interquartile Range (IQR) method, as suggested by Shang et al. (2024). The IQR, calculated as the difference between the upper (third) $Q3$ and lower (first) $Q1$ quartiles of the similarity scores, is utilized as an indicator of statistical variance. We then establish a stringent similarity threshold by selecting image tokens with similarity scores that exceed the upper bound defined as $Q3 + 1.5 \times \text{IQR}$. This approach ensures that only the most significant image tokens Z'_v, as determined by their high similarity scores, are retained.

Aggregating Unselected Tokens. Moreover, in an effort to retain the information inherent within the unselected image tokens, we calculate an average of their representations and denote it as $Z'_{v,p}$. This aggregated token is then appended to the selected tokens, a strategy that efficiently mitigates any potential loss of image information consequent to token reduction. Finally, we obtain the reduced image token sequence Z'_v.

Method	LLM	Res.	Ratio	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{C}	SEED^{I}	LLaVA	MM-Vet	AVG†
BLIP-2	Vicuna-13B	224	-	65.0	41.0	19.6	61.0	42.5	85.3	1293.8	–	–	46.4	38.1	22.4	–
InstructBLIP	Vicuna-7B	224	-	–	49.2	34.5	60.5	50.1	–	–	36.0	23.7	53.4	60.9	26.2	–
InstructBLIP	Vicuna-13B	224	-	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	–	58.2	25.6	–
IDEFICS-9B	LLaMA-7B	224	-	50.9	38.4	35.5	–	25.9	–	–	48.2	25.2	–	–	–	–
Qwen-VL-Chat	Qwen-7B	448	-	78.2	57.5	38.9	68.2	61.5	–	1487.5	60.6	56.7	58.2	–	–	–
LLaVA-1.5	Vicuna-7B	336	-	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	66.1	65.4	31.1	63.5
w. PruMerge	Vicuna-7B	336	5.5%	72.0	51.6	43.6	68.5	56.0	76.3	1350.3	60.9	50.0	50.7	45.2	21.1	55.3
w. TRIM (5%)*	Vicuna-7B	336	5%	71.5	58.4	38.4	67.9	49.1	84.8	1415.4	63.3	46.6	61.8	45.9	25.9	**57.0**
w. PruMerge+	Vicuna-7B	336	25%	76.8	56.4	42.5	68.3	57.1	84.0	1462.4	64.9	51.8	55.0	51.6	25.0	58.9
w. TRIM	Vicuna-7B	336	21%	76.4	61.4	48.1	69.1	53.7	85.3	1461.3	67.4	54.9	65.8	58.7	28.0	**61.8**
LLaVA-1.5	Vicuna-13B	336	-	80.0	63.3	53.6	71.6	61.3	85.9	1531.3	67.7	63.6	68.2	72.5	36.1	66.7
w. PruMerge	Vicuna-13B	336	5.5%	72.8	53.3	48.5	71.0	58.4	78.5	1428.2	62.3	54.5	54.4	52.4	22.0	58.3
w. TRIM (5%)*	Vicuna-13B	336	5%	75.4	56.0	50.6	70.1	50.7	85.2	1337.9	65.5	52.4	60.8	45.5	24.4	**58.6**
w. PruMerge+	Vicuna-13B	336	25%	77.8	58.9	49.7	71.0	58.6	84.4	1485.5	65.7	59.7	61.4	56.0	28.0	62.1
w. TRIM	Vicuna-13B	336	21%	75.4	59.0	53.2	72.8	54.8	86.3	1438.0	69.2	58.3	65.9	57.0	30.3	**62.8**

Table 1: **Comparison with SoTA methods on 12 benchmarks.** Res, Ratio indicate input image resolution and compression ratio of image tokens, respectively. Benchmark names are abbreviated and further detailed in Appendix B. *Select top-5% image tokens instead of automatic selection. †Average scores across 12 datasets, with MME scaled out of 2000 points. The best performance within the same ratio range is highlighted in bold.

Method	Memory (GB)	First Token (ms)	Next Token (ms)
LLaVA-1.5 (FP16)	21.8	59.1	17.6
w. PruMerge	15.1 (69.3%)	19.7 (33.3%)	17.3 (98.3%)
w. TRIM	14.9 (68.3%)	19.4 (32.8%)	17.3 (98.3%)
LLaVA-1.5 (INT8)	10.9	29.5	8.8
w. PruMerge	7.6 (69.7%)	9.9 (33.6%)	8.7 (98.9%)
w. TRIM	7.5 (68.8%)	9.7 (32.9%)	8.7 (98.9%)

Table 2: Computation Cost Analysis. Times are measured using an NVIDIA V100 GPU, representing the hardware's theoretical peak performance (batch size=1).

4 Experiment

4.1 Experiment Setup

Our experimental setup is consistent with that of LLaVA 1.5, with the key difference being that we employ our TRIM method exclusively during the instruction tuning phase. This approach ensures a fair comparison between our proposed method and the established baseline. Furthermore, we perform evaluations across 12 different datasets and compare our results with those of 5 SoTA MLLMs and one related work on token reduction. The detailed training and evaluation settings are presented in Appendix A and Appendix B.

4.2 Main Results

As shown in Table 1, after conducting experiments on 12 datasets, we found that despite reducing the image token count to 21%, our method still holds a performance level comparable to LLaVA-1.5. Moreover, it significantly outperforms previous work such as BLIP2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), IDEFICS-9B (IDEFICS, 2023), and Qwen-VL-Chat (Bai et al., 2023). Our method even surpasses LLaVA-1.5 in terms of performance on the SQA^{I} and MMB datasets. Compared to the previous work PruMerge, our method demonstrates superior performance across both token count (˜5% and ˜20%) and model size (7B and 13B), despite operating with fewer image tokens. This is particularly noticeable in the GQA, POPE and MMB datasets.

5 Analysis

5.1 Efficiency Analysis

We assessed the computational efficiency using LLMViewer analysis (Yuan et al., 2024). In a typical scenario, a 336 $\times$ 336 pixel image processed by the CLIP model yields 576 visual tokens, alongside a 40-token text prompt. After statistical analysis, PruMerge achieves a 25% compression rate, reducing the visual tokens to 144. In comparison, our method achieves a 21% compression rate, reducing the tokens to 123. Our approach significantly accelerates model inference speed and reduces memory usage, as detailed in Table 2. Notably, the time required to generate the first token is curtailed to 32.9% of the original, resulting in a significant acceleration in the inference process.

5.2 Ablation Study

We conducted an ablation study on the strategies proposed in our TRIM method, as shown in Table 3. Initially, we analyzed the automated image token selection process based on the CLIP Metric. We compared this process to a simple linear interpolation pooling, and found that our strategy can

Strategy	MMB	SEED^I
LLaVA-1.5	64.3	66.1
Random (21%)	*59.3*	*60.2*
Pooling (21%)	*61.3*	*60.6*
Automatic Selection	64.1	63.8
Aggregated Token	64.4	63.8
Training	67.4	65.8

Table 3: Impact of TRIM strategies on performance. The first row shows baseline LLaVA-1.5 results. The second and third rows illustrate the performance at 21% token sampling, achieved through random and linear interpolation respectively. The remaining rows exhibit incremental improvements from our strategies.

effectively capture key information in the image, as opposed to uniform sampling (compare the second and third rows). The usage of an additional Aggregated token to preserve sufficient image information also results in performance gains (compare the third and fourth rows). Training on the basis of the TRIM strategies can further enhance results (compare the fourth and fifth rows).

5.3 Effectiveness of TRIM Across Different Resolutions

We investigated the effectiveness of TRIM across models with varying resolutions. Following the methodology of the original LLaVA authors—who used LLaMA-2-13B-Chat (Touvron et al., 2023) and openai/clip-vit-large-patch14[1] to evaluate on LLaVA-Bench—we applied TRIM under the same conditions with a resolution of 224 × 224 pixels.

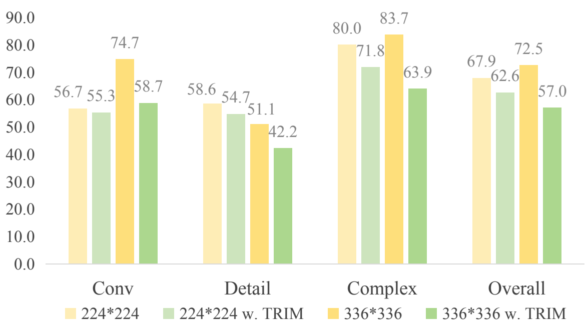

Figure 4: Comparative performance of LLaVA with and without TRIM on subsets of LLaVA-Bench across varied resolutions.

As shown in Figure 4, this approach yielded comparable performance, aligning with the results obtained using Vicuna-13B-v1.5[2] and openai/clip-vit-large-patch14-336[3] at a resolution of 336 × 336 pixels in Table 1.

5.4 Effectiveness of TRIM on Video Benchmarks

We conducted experiments on two video question-answering benchmarks: MSVD-QA (Chen and Dolan, 2011) and MSRVTT-QA (Xu et al., 2016). For each video, we extracted frames equal to the video's duration in seconds, with a maximum of seven frames due to context length limitations. We report the accuracy and average score, with assessments performed using GPT-3.5-turbo[4], as well as the first token generation time based on the average prompt length in the two datasets.

Model	**MSVD-QA**			**MSRVTT-QA**		
	Acc$_\uparrow$	Score$_\uparrow$	Time$_\downarrow$	Acc$_\uparrow$	Score$_\uparrow$	Time$_\downarrow$
LLaVA-1.5-7B	71.3	3.9	475.7	51.3	3.3	533.1
w. TRIM	67.9	3.9	76.5	49.7	3.2	84.2

Table 4: Performance comparison of LLaVA-1.5-7B with and without TRIM on the MSVD-QA and MSRVTT-QA benchmarks. Time indicates the first token generation time in milliseconds (ms).

As shown in Table 4, TRIM significantly reduces the first token generation time while maintaining comparable accuracy and average scores on both MSVD-QA and MSRVTT-QA benchmarks. This demonstrates that TRIM can effectively accelerate inference in video QA tasks without substantially compromising performance.

6 Conclusion

Our research introduced TRIM, an innovative method for reducing the image tokens in MLLMs, while maintaining performance. TRIM outperformed other methods, even with fewer tokens. Our study marks a significant step to resource-efficient MLLMs and will extend to more diverse architectures, further enhancing efficiency in the field.

Limitations

Currently, our work is primarily limited to the widely used LLaVA architecture. In the future, we aim to seamlessly integrate our method into a variety of models beyond the LLaVA architecture and into different visual encoders.

[1]https://huggingface.co/openai/clip-vit-large-patch14

[2]https://huggingface.co/lmsys/vicuna-13b-v1.5

[3]https://huggingface.co/openai/clip-vit-large-patch14-336

[4]https://platform.openai.com/docs/models#gpt-3-5-turbo

Acknowledgement

We express our gratitude to Lichao Sun for his invaluable guidance on paper structure and polish. Our appreciation is also extended to the anonymous reviewers for their constructive feedback. This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Azzedine Boukerche, Lining Zheng, and Omar Alfandi. 2020. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3):1–37.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15710–15719.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. 2023. PuMer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12890–12903, Toronto, Canada. Association for Computational Linguistics.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *Preprint*, arXiv:2402.11684.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Wentao Ge, Shunian Chen, and G Hardy Chen. 2024. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

IDEFICS. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model. `https://huggingface.co/blog/idefics`.

Yiren Jian, Tingkai Liu, Yunzhe Tao, Chunhui Zhang, Soroush Vosoughi, and Hongxia Yang. 2024. Expedited training of visual conditioned language generation via redundancy reduction. *Preprint*, arXiv:2310.03291.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. 2024. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *Preprint*, arXiv:2308.04152.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

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

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuqing Liu, Yu Wang, Lichao Sun, and Philip S Yu. 2024b. Rec-gpt4v: Multimodal recommendation with large vision-language models. *arXiv preprint arXiv:2402.08670*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. 2021. Adavit: Adaptive vision transformers for efficient image recognition. *Preprint*, arXiv:2111.15668.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

OpenAI. 2023. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Genís Prat-Ortega and Jaime de la Rocha. 2018. Selective attention: A plausible mechanism underlying confirmation bias. *Current Biology*, 28(19):R1151–R1154.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Preprint*, arXiv:2106.02034.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *Preprint*, arXiv:2403.15388.

ShareGPT. 2023. `https://sharegpt.com/`.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024a. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Dingjie Song, Sicheng Lai, Shunian Chen, Lichao Sun, and Benyou Wang. 2024b. Both text and images leaked! a systematic analysis of multimodal llm data contamination. *arXiv preprint arXiv:2411.03823*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024a. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024b. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. 2024a. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024b. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*.

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

Zhengqing Yuan, Zhaoxu Li, Weiran Huang, Yanfang Ye, and Lichao Sun. 2023. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862*.

Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, and Kurt Keutzer. 2024. Llm inference unveiled: Survey and roofline model insights. *Preprint*, arXiv:2402.16363.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. 2024. Llava-phi: Efficient multimodal assistant with small language model. *Preprint*, arXiv:2401.02330.

A Training Details

Our training method follows the instruction fine-tuning phase of LLaVA-1.5. The training data used is shown in Table 5, and the training hyperparameters are listed in Table 6. All experiments were conducted on NVIDIA A100 GPUs, with instruction-tuning for a 7B model taking approximately 8 hours and a 13B model requiring around 20 hours.

Data	Size	Response formatting prompts
LLaVA (Liu et al., 2023a)	158K	–
ShareGPT (ShareGPT, 2023)	40K	–
VQAv2 (Goyal et al., 2017)	83K	Answer the question using a single word or phrase.
GQA (Hudson and Manning, 2019)	72K	
OKVQA (Marino et al., 2019)	9K	
OCRVQA (Mishra et al., 2019)	80K	
A-OKVQA (Schwenk et al., 2022)	66K	Answer with the option's letter from the given choices directly.
TextCaps (Sidorov et al., 2020)	22K	Provide a one-sentence caption for the provided image.
RefCOCO (Kazemzadeh et al., 2014; Mao et al., 2016)	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG (Krishna et al., 2017)	86K	Provide the bounding box coordinate of the region this sentence describes.
Total	665K	

Table 5: Instruction-following Data Mixture of LLaVA-1.5-TRIM.

B Evaluation Details

B.1 Benchmark Details

The benchmarks employed in this study are detailed below.

VQA-v2 (Goyal et al., 2017); GQA (Hudson and Manning, 2019); VisWiz (Gurari et al., 2018);

Hyperparameter	Finetune
batch size	128
lr	2e-5
lr schedule	cosine decay
lr warmup ratio	0.03
weight decay	0
epoch	1
optimizer	AdamW
DeepSpeed stage	3

Table 6: **Hyperparameters** of LLaVA-1.5-TRIM are the same as the original LLaVA-1.5.

SQA$^{\mathrm{I}}$: ScienceQA-IMG (Lu et al., 2022); VQA$^{\mathrm{T}}$: TextVQA (Singh et al., 2019); POPE (Li et al., 2023c); MME (Perception) (Fu et al., 2023); MMB: MMBench (Liu et al., 2023b); SEED$^{\mathrm{I}}$: SEED-Bench-IMG (Li et al., 2023a); LLaVA$^{\mathrm{W}}$: LLaVA-Bench (In-the-Wild) (Liu et al., 2023a); MM-Vet (Yu et al., 2023);

- VQA-v2 (Goyal et al., 2017) is a dataset for VQA containing 265,016 images with at least 3 questions per image and 10 ground truth answers per question. *Accuracy* is used as the metric.
- VisWiz (Gurari et al., 2018) is a VQA dataset designed for assisting blind people. Each image in the dataset is accompanied by a spoken question and 10 crowdsourced answers. The challenge of the dataset includes predicting the answer to a visual question and whether a question can be answered. *Accuracy* is used as the metric.
- POPE (Li et al., 2023c) is a benchmark designed to evaluate the object hallucination issue in MLLMs. The evaluation metric is the *F1* score.
- GQA (Hudson and Manning, 2019) consists of 12,578 questions for real-world reasoning and compositional question answering. *Accuracy* is used as the metric.
- MME (Fu et al., 2023) is a benchmark with 2,374 questions spanning 14 subtasks. *Accuracy* is used as the metric.
- TextVQA (Singh et al., 2019) comprises 5,000 questions and *Accuracy* is used as the metric.
- MM-Vet (Yu et al., 2023) comprises 218 questions, each requiring multiple capabilities to solve and provided with multiple groundtruths for a flexible match. *Accuracy* is adopted as the metric.
- ScienceQA (Lu et al., 2022) contains 4,201 questions encompassing different subjects and categories. *Accuracy* is adopted as the metric.
- LLaVA-Bench (In-the-Wild) (Liu et al., 2023a) contains 60 open-ended questions and uses text-based GPT-4 (OpenAI et al., 2024) as a judge to score answers in a pairwise fashion. *Score Ratio* between candidate answers and anchor answers from GPT-4 is adopted as the metric.
- MMBench (Liu et al., 2023b) (dev set) consists of 4,329 multiple-choice questions across 20 ability dimensions, using *Accuracy* under circular evaluation as the metric.
- SEED-Bench (Li et al., 2023a) (image set) comprises 14,233 multiple-choice questions across 9 dimensions. *Accuracy* is adopted as the metric.

B.2 Evaluation Prompts

We standardize the prompt formats used for evaluation according to LLaVA-1.5, as presented in Table 7.

Data	Response formatting prompts
LLaVA-Bench, MM-Vet	–
VQAv2, GQA, TextVQA, MME, POPE	Answer the question using a single word or phrase.
ScienceQA, MMBench, SEED-Bench	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 7: Response format prompt for evaluation.