InfiMM: Advancing Multimodal Understanding with an Open-Sourced Visual Language Model

Haogeng Liu[1,2], **Quanzeng You**[3], **Yiqi Wang**[3], **Xiaotian Han**[3], **Bohan Zhai**[3], **Yongfei Liu**[3]
Wentao Chen[3], **Yiren Jian**[3], **Yunzhe Tao**[3], **Jianbo Yuan**[3], **Ran He**[1,2], **Hongxia Yang**[3]*
[1]New Laboratory of Pattern Recognition (NLPR), Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]ByteDance, Inc
liuhaogeng2022@ia.ac.cn, rhe@nlpr.ia.ac.cn
{quanzeng.you, xiaotian.han, hongxia.yang}@bytedance.com

Abstract

In this work, we present InfiMM, an advanced Multimodal Large Language Model that adapts to intricate vision-language tasks. InfiMM, inspired by the Flamingo architecture, distinguishes itself through the utilization of large-scale training data, three-stage training strategies, and diverse large language models. This approach ensures the preservation of Flamingo's foundational strengths while simultaneously introducing augmented capabilities. Empirical evaluations across a variety of benchmarks underscore InfiMM's remarkable capability in multimodal understanding. The code and model can be found at: https://huggingface.co/Infi-MM.

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) have shown a transformative evolution through the integration of pretrained vision encoders with Large Language Models (LLMs). Seminal contributions to this domain include Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023b), BLIP-2 (Li et al., 2023), and MiniGPT-4 (Zhu et al., 2023a). MLLMs demonstrate exceptional proficiency across a variety of tasks, including image captioning, visual question answering, and more complex activities such as generating code from images, converting image plots into Markdown format tables, and simulating web browsing.

For effective integration of pretrained vision encoders with large language models, careful design of vision-language connector modules is essential. These modules play a critical role in transforming and aligning visual tokens to formats compatible with Large Language Models, as well as effectively leveraging these tokens. Models like Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) utilize Perceiver Resampler/Q-Former techniques, offering flexibility and nuanced alignment with language counterparts. However, this approach can incur high computational costs and risk information loss (Cha et al., 2023). Conversely, models such as LLaVA and MiniGPT-v2 employ simpler Multi-Layer Perception (MLP) strategies, reducing computational complexity but potentially sacrificing nuanced representation of visual data.

In the utilization of the transformed tokens, architectures akin to Flamingo employ cross-attention mechanisms, enabling nuanced interactions between token types without necessitating an expansion of the token sequence length. This method effectively manages computational load. Conversely, LLaVA-style models adopt a direct concatenation approach, which, while straightforward, leads to an augmentation in token sequence length and computational complexity.

Though efficient in the inference stage, few works adopt the Flamingo-style architecture. OpenFlamingo (Awadalla et al., 2023) and IDEFICS (Laurençon et al., 2023a) are two reproductions of the Flamingo. However, as they use less capable language models and limited training data, their performance could be improved. Further we propose three-stage training strategies for vision-language alignment, vqa knowleage injection and unreshing conversation abality. We utilize a stronger vision encoder, language model, and higher-quality data to build a stronger model. We anticipate that this will foster development within the field.

2 Related Work

Large language models (LLMs) have made significant advancements (OpenAI, 2023; Chowdhery et al., 2022; Bai et al., 2022; Touvron et al., 2023; Tunstall et al., 2023). These models are powerful in chatting and can finish many tasks only with different instructions. Though impressive, these

* Corresponding Authors

Findings of the Association for Computational Linguistics ACL 2024, pages 485–492

models are limited to only the language domain but not other modalities. They have subsequently been extensively utilized in multimodal tasks such as image-to-text generation and video-to-text generation (Zhang et al., 2023a; Xu et al., 2023; Huang et al., 2023; Alayrac et al., 2022; Wang et al., 2022; Li et al., 2023; Liu et al., 2023b), giving rise to a new class of models called multimodal large language model (MLLM).

Flamingo (Alayrac et al., 2022) leverages pretrained language models within the MLLM framework, employing gated-cross attention to integrate visual information into textual sequences. In contrast, BLIP-2 (Li et al., 2023), MiniGPT4 (Zhu et al., 2023a), and LLaVA (Liu et al., 2023b) propose a novel approach by converting visual signals into soft tokens and directly integrating them into language models. Utilizing gated-cross attention as a modality connector introduces more trainable parameters but can potentially reduce inference cost as the visual signal will not be turned into soft tokens, thereby not increasing the sequence length of large language models.

While numerous open-source projects have emerged following the architectures of LLaVA and BLIP-2, there needs to be more emphasis on the Flamingo-style architecture. OpenFlamingo (Awadalla et al., 2023) and IDEFICS (Laurençon et al., 2023a) represent two open-source models adopting the Flamingo-style approach. However, due to constraints imposed by their language model and vision encoder, their capabilities could be more remarkable.

In this study, we adopt the Flamingo framework and harness a more potent combination of language model and vision encoder to construct a robust model. Additionally, we employ higher-quality data for training, aiming to enhance the model's strength. These efforts will likely result in a more formidable model and contribute to the advancement of research in MLLMs.

3 Method

3.1 Model Architecture

We show our model architecture in Figure 1. InfiMM is inspried by Flamingo (Alayrac et al., 2022). The details of our model will be discussed in the following :

Large Language Model: InfiMM reveals the impact of LLMs with different scales and architectures. For the 7B setting, InfiMM adapts pretrained Zephyr as a language model. For the 13B setting, InfiMM adapts either LLaMA2 (Touvron et al., 2023) or its finetuned version Vicuna (Chiang et al., 2023) as the language model.

Vision Encoder: InfiMM utilizes the EVA2-CLIP-G (Sun et al., 2023) as default vision encoder, which fixes the input resolution to 224×224.

Connector: InfiMM adapts the Perceiver Resampler and Gated Cross-attention as the V-L connectors. Percevier Resampler consists of cross-attention layers and learnable queries. This could compress vision features to fixed 32 vision tokens. Meanwhile, Gated Cross-attention layers are used for vision-language interaction.

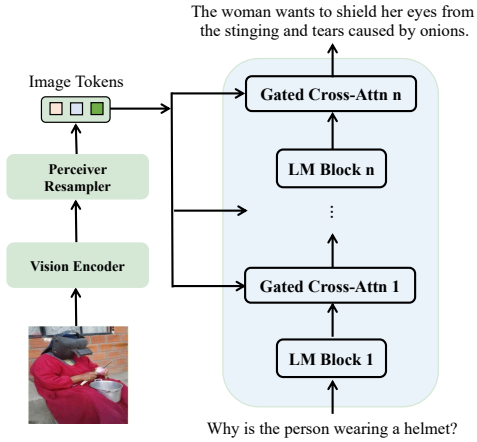

Figure 1: The overview architecture of InfiMM. InfiMM consists of a vision encoder, a Perceiver Resampler, and a large language model with a Gated Cross-attention module.

3.2 Training details

We have established a three-stage training procedure for improving InfiMM's overall ability, as shown in Figure 2. These stages are denoted as Pretraining (PT), Multi-Task Training (MTT), and Instruction Finetuning (IFT). The PT stage aims to align vision-language modalities, MTT stage integrates vision-language question-answering knowledge, and IFT stage significantly improves the model's conversational abilities.

Pretraining Stage: This stage focuses on the initial alignment of vision features and language features. During this stage, both the vision encoder and large language model are frozen, with only the Gated Cross-attention module and the Perceiver Resampler being learnable. The training dataset involves a diverse set of image-text pairs (LAION (Schuhmann et al., 2022),

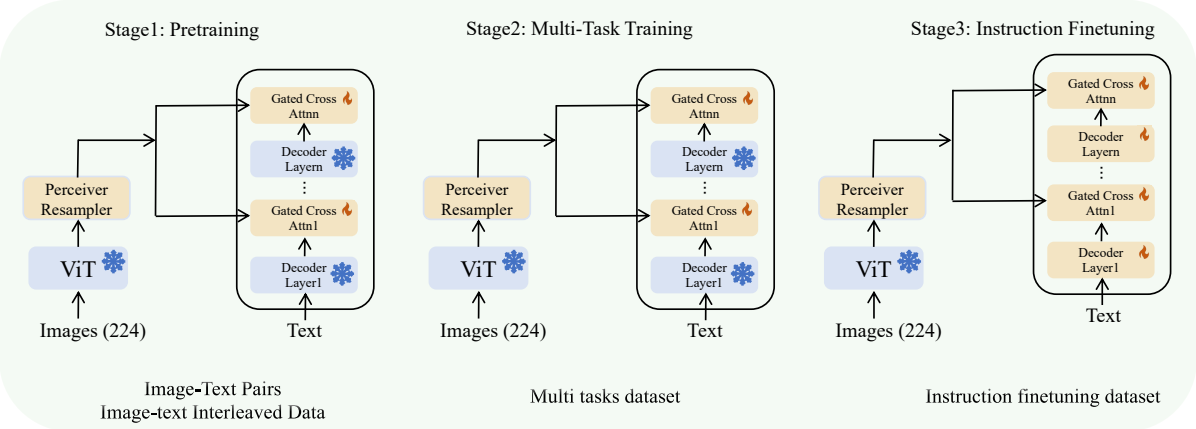

Figure 2: The training pipeline of InfiMM. The language model is trainable only in the Instruction Finetuning Stage. And ViT is frozen during the entire training process.

Table 1: Details on the training data of Pretraining Stage.

Dataset	Type of Data	Samples
OBELICS	Unstructured Web Docs	101M
MMC4	Unstructured Web Docs	53M
LAION	Image-Text Pairs	115M
COYO	Image-Text Pairs	238M
LAION-COCO	Image-Text Pairs	140M
PMD	Image-Text Pairs	20M
Total	-	667M

Table 2: Details on the training data of Multi-Task Training Stage.

Task	Dataset	Samples
Image Caption	COCO Caption (Chen et al., 2015)	410k
	TextCaps (Sidorov et al., 2020)	110k
	VizWiz Caption (Gurari et al., 2020)	110k
General VQA	VQAV2 (Antol et al., 2015)	443k
	OKVQA (Marino et al., 2019)	9k
	VizWiz VQA (Gurari et al., 2018)	20k
	GQA (Hudson and Manning, 2019)	471k
	A-OKQA (Schwenk et al., 2022)	17k
Text-oriented VQA	TextVQA (Singh et al., 2019)	34k
	OCRVQA (Mishra et al., 2019)	166k
	STVQA (Biten et al., 2019)	26k
	DocVQA (Mathew et al., 2021)	63k
	LLaVAR (Zhang et al., 2023b)	16k
Total	-	1.86M

COYO (Byeon et al., 2022), LAION-COCO, PMD (Singh et al., 2022)) and unstructured multimodal web documents(OBELICS (Laurençon et al., 2023a), MMC4 (Zhu et al., 2023b)), all sourced from public domains and PMD is only used in 13B LLMs. We have also filtered out low-quality data, resulting in the following dataset utilized:

Multi-Task Training Stage (MTT) As the dataset used in pretraining is mainly instance-level alignment and has a lot of noise, we introduce Multi-Task Training for higher-quality knowledge injection. In this stage, we focus on supervised training on different tasks, including image captions and visual question-answering. Also include other domains, like scene-based datasets (Chen et al., 2015), (Hudson and Manning, 2019), and OCR based datasets (Sidorov et al., 2020), (Mishra et al., 2019), (Zhang et al., 2023b) etc. We keep the trainable parameters same with the first stage. Detailed information about training datasets is listed in Table 2.

Instruction Finetuning Stage (IFT): In this final stage, our goal is to make the model better follow user instructions and develop the "chat" version of InfiMM. We only keep the ViT frozen while all the other parameters are trainable. In this stage, we utilize the LLaVA-665k (Liu et al., 2023a) instruction finetuning dataset for training.

4 Experiment

We evaluate InfiMM across a diverse array of tasks. For image caption, we test our model with COCO and Flickr30k. For general VQA tasks, we leverage benchmarks such as OKVQA (Marino et al., 2019), VQAV2 (Antol et al., 2015) and TextVQA (Singh et al., 2019). On these dataset, we only evaluate the pretrained model in a zero-shot and few-shots manner. Results can be found in Table 4.

We also assess the logical reasoning capabilities of our model by employing newly introduced benchmarks, including MM-VET (Yu et al., 2023), MME (Fu et al., 2023), MMbench (Liu et al., 2023c), InfiMM-Eval (Han et al., 2023), and MMMU (Yue et al., 2023). Notably, the MMMU (Yue et al., 2023) presents challenging tasks that de-

Table 3: Results of InfiMM-Chat on general VQA task.

Model	ScienceQA-Img	MME	MMVet	InfiMM-Eval	MMbench	MMMU-Val	MMMU-test
Otter-9B	-	1292/306	24.6	32.2	-	22.69	-
IDEFICS-9B-Instruct	60.6	-/-	-	-	-	24.53	-
LLaVA-1.5	71.6	1531/295	35.4	32.6	67.7	36.4	33.6
QWen-VL-Chat	68.2	1488/361	-	37.4	60.6	35.9	32.9
InfiMM-Zephyr-7B-Chat	71.1	1406/327	32.8	36.0	59.7	39.4	35.5
InfiMM-Vicuna-13B-Chat	74.0	1461/324	36.0	40.0	66.7	37.6	34.6
InfiMM-Llama-13b-Chat	73.0	1445/338	39.2	41.4	66.4	39.1	35.2

Table 4: Results on general VQA task. Here we report zero-shot and four-shots result of InfiMM.

Model	Shots	COCO	Flickr30k	OKVQA	VQAv2	TextVQA
IDEFICS-9B	0	46	27.3	38.4	50.9	25.9
	4	93	59.7	45.5	55.4	27.6
IDEFICS-80B	0	91.8	53.7	45.2	60	30.9
	4	110.3	73.7	52.4	64.6	34.4
InfiMM-Zephyr-7B	0	78.8	60.7	17.1	33.7	15.2
	4	108.6	71.9	50.5	59.1	34.3
InfiMM-Vicuna13B	0	69.6	49.6	49.2	60.4	32.8
	4	118.1	81.4	53.7	64.2	38.4
InfiMM-Llama2-13B	0	85.4	54.6	26.4	51.6	24.2
	4	125.2	87.1	55.5	66.1	38.2

mand advanced subject knowledge and deliberate reasoning at a collegiate level. These tasks span diverse fields such as physics, chemistry, and biology. The MM-VET benchmark assesses the integrated capabilities of models.

4.1 In-context Learning Ability

We conduct a comparative analysis of InfiMM's capacity for in-context learning against that of IDEFICS (Laurençon et al., 2023b) in Table 4, which represents the original leading Flamingo-style architecture model under zero or four-shot conditions. Our findings reveal that InfiMM outperforms IDEFICS across all benchmark metrics. Notably, even our 13B model demonstrates superiority over IDEFICS' 80B model, underscoring the efficacy of our training methodology.

4.2 General Logical Reasoning Benchmarks

In Table 3, we compare our method with various methods. InfiMM shows competitive performance on both benchmarks, especially the MMMU benchmark, which needs complicated vision and language understanding capability. We are superior to most of the previous models in both the validation dataset and testing dataset.

4.3 Influence of Training Stage

Analysis of Table 5 reveals that the introduction of MTT significantly enhances the model's knowledge assimilation, which shows a significant improvement in VQA tasks like OKVQA, VQAv2, TextVQA, and GQA (Hudson and Manning, 2019). However, constrained by the limited response format inherent in MTT, the model exhibits subpar performance in open-ended tasks (as observed in InfiMM-Eval). Notably, following Instruction Finetuning, the model demonstrates improved proficiency in handling more diverse and flexible tasks.

Table 5: Ablation study on training stages. MTT and IFT mean Multi-Task Training and Instruction Finetuning Stage.

Model	MTT	IFT	OKVQA	VQAv2	TextVQA	GQA	MMMU	InfiMM-Eval
InfiMM-Llama13B		✓	61.2	75.0	41.3	57.9	37.1	36.3
InfiMM-Llama13B	✓		63.4	76.9	45.0	62.0	36.1	28.3
InfiMM-Llama13B-Chat	✓	✓	62.3	78.5	44.6	61.2	39.2	41.4

5 Limitations

Although InfiMM demonstrates robust performance in vision-language modeling while maintaining a balanced computational load for processing multiple images, its efficacy is hampered by the constraint of limited image size, thereby restricting its ability to address complex visual content effectively.

6 Conclusion

In this study, we introduce InfiMM, an advanced multimodal large language model that significantly advances the field of visual language understanding. InfiMM's architecture, inspired by Flamingo and enhanced by our methodological innovations, demonstrates a delicate balance between computational efficiency and the capacity to handle nuanced visual-language tasks. Evaluation on various benchmarks highlights InfiMM's remarkable ability to understand complex scenes and shows good reasoning ability. InfiMM represents a significant step forward in the multimodal understanding domain.

Ethics Statement

Our MLLMs are constructed upon pre-trained LLMs. Consequently, our models inherit the potential risks associated with LLMs, such as the generation of biased, inappropriate, discriminatory, offensive, misleading, or even harmful contents.

Additionally, our models undergo training on publicly accessible datasets including LAION, COYO, LAION-COCO, PMD, MMC4, and others. Despite the extensive usage of these datasets, the presence of discriminatory, biased, or sensitive content cannot be ruled out. Given that our models inherently assimilate such information during the training process, prudence is warranted in their application.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, et al. 2022. Constitutional ai: Harmlessness from ai feedback.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, et al. 2022. Palm: Scaling language modeling with pathways.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

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.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer.

Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, Heng Wang, and Hongxia Yang. 2023. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, et al. 2023. Language is not all you need: Aligning perception with language models.

Drew A Hudson and Christopher D Manning. 2019. 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.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023a. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023b. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. Mmbench: Is your multi-modal model an all-around player?

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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-4 technical report.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

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.

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, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

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.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Peng Wang, An Yang, Rui Men, Junyang Lin, et al. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu:

A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023a. Transfer visual prompt generator across llms.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. Llavar: Enhanced visual instruction tuning for text-rich image understanding.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Summary of Evaluation Benchmarks

We provided a detailed summary of evaluation benchmarks we used and their corresponding metrics in Table 6.

Table 6: Details on the test dataset.

Task	Dataset	Description	Split	Metric
General VQA	VQAV2	VQA on natural images	test-dev	VQA Score(↑)
	OKVQA	VQA on natural images but need world knowledge	val	VQA Score(↑)
	GQA	VQA on scene understandig and reasoning	test-dev	EM(↑)
	TextVQA	VQA about text in natural scene	val	VQA Score(↑)
Other Benchmarks	MME	Evalutaion for MLLM on perception and cognition	Perception and Cognition	Accuracy(↑)
	MM-VET	Dialog style VQA on integrated ability	test	GPT-4 score(↑)
	MMbench	Comprehensive evalutaion with multi choice VQA	test	Accuracy(↑) score(↑)
	InfiMM-Eval	Complex Open-ended Reasoning	test	GPT-4 score(↑)
	MMMU	College-level multi choice VQA	val and test	Accuracy(↑)

A.1 Training Configuration

Table 7: Details of the training Configuration.

Configuration	Pretraining	Multi-Task Training	Instruction Finetuning
ViT init.	EVA2-CLIP2-g	EVA2-CLIP2-g	EVA2-CLIP2-g
LLM init.	LLaMA2-13b	LLaMA2-13b	LLaMA2-13b
Gated Cross-attention init.	random	InfiMM 1st stage	InfiMM 2rd stage
Image resolution	224		
ViT sequence length	257		
Perceiver Resampler length	64		
LLM sequence length	32 (IT);384 (IIT)	128	512
Optimizer	AdamW		
Optimizer hyperparameter	$\beta_1 = 0.9, \beta_2 = 0.95, eps = 1e^{-8}$	$\beta_1 = 0.9, \beta_2 = 0.999, eps = 1e^{-5}$	
Peak learning rate	$1e^{-4}$	$1e^{-5}$	$5e^{-6}$
Minimum learning rate	$1e^{-4}$	$1e^{-6}$	$5e^{-7}$
Learning rate schedule	cosine decay		
Weight decay	0.1		
Gradient clip	1.0		
Training steps	285k	10k	6k
warm steps	6k	500	300
Global batch size	5120	256	64
Gradient accumulation steps	2	1	1
Gradient ACC.	2	1	2
Numerical precision	bfloat16		
Gradient checkpointing	×	✓	×
Deepspeed Zero Stage	2		
Training resource	40 NVIDIA A100-SXM-80GB	32 NVIDIA A100-SXM-80GB	