ChiImageQA Technical Report

Yuhui Zheng
City University of Hong Kong
cs.yuhuizheng@my.cityu.edu.hk

Abstract

Recent work on MiniGPT-4 has demonstrated that a ViT and Q-former structure for BLIP2 can be properly aligned to large language models (LLMs) by simply training a linear layer to acquire numerous advanced multimodal capabilities, such as generating detailed image descriptions and creating websites from hand-drawn sketches. In this era of rapid updating and iteration of current state-of-the-art (SOTA) large language models, this method of constructing large-scale vision-language models (LVLMs) is extremely efficient. However, due to the lack of Chinese data for high-quality instruction fine-tuning and low-quality image-text alignment in MiniGPT-4, its response quality to Chinese instructions is relatively low. To ameliorate this issue, we introduce ChiImageQA, a LVLM based on the architecture of MiniGPT-4, which utilises 18.8k Chinese and English instruction fine-tuning data to fine-tune only the linear layer of MiniGPT-4. To ensure that ChiImageQA has good scalability for multi-image reasoning tasks in the future, while minimizing computational overhead, we implement a efficient multimodel integration approach. Compared to MiniGPT-4, ChiImageQA demonstrates responses to Chinese instructions that are more comprehensive, logical, and well-structured and shows a significant improvement. Our code, linear layer checkpoint, instruction dataset are available at https://github.com/cszhengyh/ImageQA.

1. Introduction

In recent years, large language models (LLMs) [4, 10, 22, 24, 29, 34] have attracted wide attention due to their powerful capabilities in text generation and comprehension. These models can perform a variety of intricate linguistic tasks in a zero-shot manner and be further aligned with user intent through fine-tuning instructions, showcasing strong interactive capabilities and the potential to enhance productivity as intelligent assistants. However, native LLMs only live in the pure-text world, lacking the ability to handle other commonmodalities (such as images, speech, and videos), resulting in great restrictions on their application scope. Notably, GPT-4 [23], a large-scale multimodal model, has been recently introduced and demonstrated several impressive capabilities of vision-language understanding and generation, for example, GPT-4 can produce detailed and accurate image descriptions, explain unusual visual phenomena, and even construct websites based on handwritten text instructions, which are closed-source with their inner mechanisms opaque. This limits the democratization of AI technologies and the scope of community-driven innovation and development.

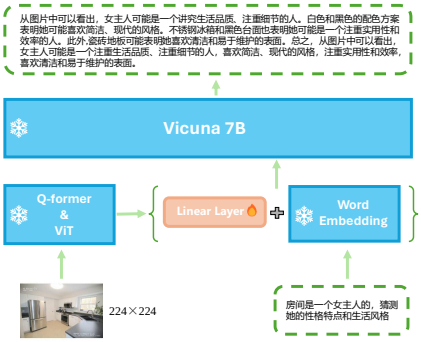

Figure 1. **The architecture of ChiImageQA**. It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. ChiImageQA only requires fine-tuning the linear projection layer to align the visual features with the Vicuna.

To unlock the secrets behind GPT-4, a group of open-source Large Vision Language Models (LVLMs) [1, 3, 7, 8, 12, 17–20, 23, 25, 31, 41, 43, 47] have been developed to enhance LLMs with the ability to perceive and understand visual signals. These large-scale vision-language models demonstrate promising potential in solving real-world vision-central problems. Nevertheless, despite that lots of works have been conducted to explore the limitation and potency of LVLMs, current open-source LVLMs always suffer from low quality and hallucination when processing Chinese queries, which hinders further exploration and application of LVLMs in the Chinese context. We believe that the issues encountered when handling Chinese

(a) MiniGPT-4 (b) ChiImageQA

Figure 2. Model generations from MiniGPT-4 and ChiImageQA.

queries may stem from the absence of Chinese data in the model training dataset.

To substantiate our hypothesis, we explore a way out and present a Chinese vision-language model named ChiImageQA. As illustrated in Fig. 1, ChiImageQA utilizes an advanced LLM, Vicuna [9], which is built upon LLaMA [34]. In terms of visual perception, we employ the same pre-trained vision components of BLIP-2 [19] that consists of a ViT-G/14 from EVA-CLIP [15] and a Q-Former network. ChiImageQA trains a single projection layer to align the encoded visual features with the Vicuna language model and freezes all the other vision and language components. The work of InstructionGPT-4[37] demonstrates that it only fine-tuned with 200 instructions consistently outperforms the original MiniGPT-4 in various popular benchmarks such as MME, MMBench and VQA datasets. Inspired by this, we utilize high-quality single-turn Chinese and English instruction data[2] to fine-tune the linear layer of ChiImageQA on Colab[1], aligning visual features with the Vicuna language model, thereby enhancing the naturalness of generated language for Chinese queries and its usability in Chinese context. Moreover, considering potential future research tasks that may involve endowing ChiImageQA with the ability to perform multi-image reasoning tasks, we implemented an efficient multimodal integration method to achieve better performance without increasing computational overhead.

Through multiple Chinese-English interaction with ChiImageQA, we found that ChiImage QA is capable of perceiving and understanding visual inputs, generating desired responses according to given prompts, and accomplishing various vision-language tasks such as image captioning, question answering, text-oriented question answering, and visual grounding. Moreover, compared to MiniGPT-4, ChiImageQA performs better on multiple visual language tasks of MiniGPT-4, such as generating detailed cooking recipes from food photos, writing stories or poems inspired by images, writing advertisements for products in images, identifying problems shown in photos and providing corresponding solutions, and retrieving rich facts about people, movies, or art directly from images, creating websites based on handwritten text instructions, explaining unusual visual phenomena, among other capabilities. Especially noteworthy is the significant improvement in response to Chinese instructions. This further demonstrates that for models based on the structure of MiniGPT-4, only fine-tuning with instruction data can effectively align visual information with LLMs and incorporating Chinese training data into the instruction dataset enhances the model's responsiveness to Chinese instructions.

[1] https://colab.research.google.com/

Vision Encoder	Linear Layer	LLM	Total
1.178B	0.009B	7B	8.197B

Table 1. Details of ChiImageQA model parameters.

We present a summary of our key findings:

- Our research provides a solution for achieving optimal responses to Chinese and English instructions in visual language tasks under computational resource constraints[1].
- Our findings suggest that fine-tuning the models based on the structure of MiniGPT-4 with Chinese instructions proves to be an effective method for enhancing its response quality to Chinese instruction in visual language tasks.

2. Related Work

2.1. Large Language Model

Tremendous success has been experienced by large language models in recent years, attributed to the expansion of training data and the increase in parameter count. Foundational progress was laid by early models like BERT [13], GPT-2 [27], and T5 [28]. GPT-3 [4], boasting a massive scale of 175 billion parameters, marked a significant breakthrough across numerous language benchmarks. This advancement sparked the development of various other large language models, including Megatron-Turing NLG [30], Chinchilla [16], PaLM [10], OPT [44], BLOOM [29], and LLaMA [34], among others. Several emergent abilities, exclusive to large models, were further discovered by Wei et al. [36], emphasizing the significance of scaling up in large language model development. InstructGPT [24] and ChatGPT [22] enable conversational interactions with humans and can address a wide range of diverse and complex questions by aligning the pre-trained large language model GPT-3 with human intent, instructions, and human feedback. More recently, models like Alpaca [33] and Vicuna [9] have been developed based on LLaMA [34] and

demonstrate similar performance.

2.2. Large-scale Vision Language Model

With the rapid advancement of LLMs [2, 23, 26], researchers have commenced the development of more potent LVLMs based on LLMs [1, 7, 8, 12, 17, 18, 18–20, 25, 31, 41, 43, 47]. This approach use autoregressive language models as decoders in vision-language tasks, which takes advantage of cross-modal transfer, allowing knowledge to be shared between language and multimodal domains. Pioneering studies like VisualGPT [5] and Frozen [35] have demonstrated the benefits of employing a pre-trained language model as a vision-language model decoder. Flamingo [1] was then developed to align a pre-trained vision encoder and language model using gated cross-attention, and was trained on billions of image-text pairs, showcasing impressive in-context few-shot learning capabilities. Q-Former, proposed by BLIP-2 [19], efficiently aligns the frozen vision foundation models and Flan-T5[11]. LLAVA [20] and MiniGPT-4 [47] introduce visual instruction tuning to enhance instruction following capabilities in LVLMs. Additionally, mPLUG-DocOwl [41] integrates document understanding capabilities into LVLMs by introducing digital documents data. Kosmos2 [25], Shikra [7], and BuboGPT [45] further augment LVLMs with visual grounding abilities, facilitating region description and localization. Qwen-VL [3] integrates image captioning, visual question answering, OCR, document understanding, and visual grounding capabilities. The resulting model achieves outstanding performance on these diverse style tasks. Most recently, PaLM-E[14], featuring 562 billion parameters, has been developed to integrate real-world continuous sensor modalities into an LLM, thereby establishing a connection between real-world perceptions and human languages. GPT-4[23] has also been recently released, showcasing more powerful visual understanding and reasoning abilities after pre-training on a vast collection of aligned image-text data.

2.3. Separating LLM and Vision Model

Unlike LVLM, another line of work has abandoned the alignment of LLMs and vision foundation models. Instead, it leverages LLM's powerful capabilities in using tools, coding, multi-turn dialogue, high-quality Q&A, etc., to collaborate with vision model to solve different visual language tasks. For instance, Visual ChatGPT [38] and MM-REACT [40] showcase how ChatGPT can act as a coordinator, integrating with diverse visual foundation models and facilitating their collaboration to tackle more complex challenges. ChatCaptioner [46] treats ChatGPT as a questioner, prompting diverse questions for BLIP-2 to answer. Through multi-round conversations, ChatGPT extracts visual information from BLIP-2 and effectively summarizes the image content. Video ChatCaptioner [6] extends this approach, applying it to video spatiotemporal understanding. ViperGPT [32] demonstrates the potential of combining an LLM with different vision models to address complex visual queries programmatically.

3. Method

3.1. Model Architecture

ChiImageQA utilizes the Vicuna [9] as our language decoder, which is constructed upon LLaMA [34] and can perform a wide range of complex linguistic tasks. For visual perception, we employ the same visual encoder as used in BLIP-2 [19], a ViT backbone [15] coupled with their pre-trained Q-Former. Both language and vision models are open-sourced. A linear projection layer is used to bridge the gap between the visual encoder and LLM, with an overview of our model displayed in Fig. 1. Tab. 1 illustrates the parameter sizes of various components of the ChiImageQA.

3.2. Multimodal Integration Approach

In visual language tasks, models encode images and text to obtain image embeddings and text embeddings. These embeddings are concatenated and inputted into LLMs to generate responses. This concatenation process is referred to as multimodal integration. Different LVLMs offer different integration strategies. In BLIP2, the integration strategy involves placing the image embedding consistently before the text embedding, limiting the flexibility to insert it at arbitrary positions within the text embedding, which may not be ideal for scenarios involving multi-image reasoning. In the work of Qwen-VL [3], two special tokens, <img> and </img>, are introduced to locate the position of the image embedding, allowing for flexible integration. However, this approach introduces special tokens that are not present in Qwen's vocabulary, necessitating the retraining of the word embedding layer and LM head layer to learn these special tokens' embedding and meaning, thereby increasing the computational overhead.

Considering these issues, to ensure that ChiImageQA has good scalability for multi-image reasoning tasks in the future and support the insertion of image embedding at arbitrary positions within the text embedding without increasing computational overhead, we implement a efficient multimodal integration approach, which uses placeholders from the vocabulary of LLMs to temporarily replace the image embedding. Since the output length of Q-former is fixed at 32 tokens, the number of placeholders in the text is 32. After encoding the image and text, the placeholders in the text embedding are replaced with the image embedding before inputting them into the LLMs. Additionally, for aesthetic purposes in ChiImageQA, we use the token <ImageHere> in the query to indicate the position where the image

Chinese	English	Total
9,527	9,351	18,878

Table 2. Details of MiniGPT-4 training dataset from MMPretrain.

is inserted. Before inputting the word embedding layer, we replace <ImageHere> in the query with 32 placeholders. To further enhance the LLMs' understanding of multimodal features, we use the non-special token <Img> and </Img> to hint the LLMs that this position represents an image. For example, 图像1 <Img><ImageHere></Img> 和图像2 <Img><ImageHere></Img>中的不同之处是什么?

3.3. Instruction Fine-tuning

During the finetuning, we use the MiniGPT-4's second-stage fine-tuning predefined prompts to refactor our input for producing more natural and reliable language outputs.

4. Experiment

4.1. Dataset

We perform instruction fine-tuning on the MiniGPT-4 training dataset from MMPretrain[2]. The images in this dataset have a pixel size of 224×224. As shown in Tab. 2, these image instructions consist of instructions in both Chinese and English, where the Chinese instructions is obtained by translating English instructions through ChatGPT. In order to facilitate training, we reprocessed annotations for the dataset[3].

4.2. Implementation Details

Our ChiImageQA model utilizes Vicuna [9] 7B checkpoint as language model, the BLIP2 [19] pre-trained model for Q-former and the ViT-G/14 checkpoint from EVA-CLIP [15] for vision encoder. We train ChiImageQA for 10 epoches with AdamW [21] optimization on Colab [1]. The batch size is set to 16. We choose 1×10^{-3} as the peak learning rate and set the maximum length of input tokens to 100. The warmup strategy is used to adjust the learning rate during training, the warmup steps is set to be 500. We utilizes loss scaling is used to avoid underflow.

5. Conclusion and Future Work

We introduced ChiImageQA, a LVLM based on the MiniGPT-4 architecture, aimed at enhancing the understanding of Chinese context in visual language tasks. ChiImageQA surpasses MiniGPT4 in response quality to both Chinese and English instructions across various visual language tasks, while requiring minimal computational overhead. In the future, with unrestricted computational resources, we can further enhance ChiImageQA's capabilities in the following key dimensions:

5.1. Chinese Instruction Response Capability

- **Larger Chinese Language Model.** In the original MiniGPT-4 architecture, Vicuna and LLaMA are used as language models, but their support for Chinese is relatively weak, resulting in poor support for Chinese instructions in the trained LVLM. However, there are now many open-source Chinese LLMs available, such as Qwen [26], BaiChuan [39], GLM [42], etc. By leveraging larger-scale Chinese LLMs as language models, we will further enhance the response quality to Chinese instructions.
- **More Data and Training.** Additionally, when our computational resources are unlimited, we can pre-train the linear layer using a large amount of crowdsourced aligned image-text data. Simultaneously, we can leverage more high-quality Chinese and English instruction data to fine-tune the LLM, linear layer, and Q-former. To ensure that the visual feature extraction capability remains uncompromised, we typically choose to freeze the visual model. Training more parameters with the extensive data mentioned above would lead to an improvement in the response quality to Chinese instructions.
- **Better Vison Encoder.** Certainly, a better visual model would also contribute to improving the response quality to Chinese instructions. It seems that Large Vision Models might indeed be necessary.

5.2. Other capabilities

By leveraging state-of-the-art LVLMs such as GPT-4 and constructing training data for different tasks such as multi-image reasoning and multi-turn dialogue, we can train ChiImageQA to gain additional capabilities.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 3

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv:2305.10403*, 2023. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3

[2] https://huggingface.co/datasets/deepHug/minigpt4_training_for_MMPretrain

[3] https://github.com/cszhengyh/ImageQA/blob/main/instruction_data.zip

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[5] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3

[6] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023. 3

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 1, 3

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv:2209.06794*, 2022. 1, 3

[9] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 3, 4

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 2

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023. 1, 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 2, 3, 4

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045*, 2023. 1, 3

[18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv:2305.03726*, 2023. 3

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023. 2, 3, 4

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023. 1, 3

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[22] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. 1, 2

[23] OpenAI. Gpt-4 technical report, 2023. 1, 3

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 1, 2

[25] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023. 1, 3

[26] Qwen. Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts), 2023. 3, 4

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[29] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1, 2

[30] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 2

[31] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun

Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv:2307.05222*, 2023. 1, 3

[32] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 3

[33] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 2

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2, 3

[35] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[36] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification. 2

[37] Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*, 2023. 2

[38] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[39] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 4

[40] Zhengyuan Yang*, Linjie Li*, Jianfeng Wang*, Kevin Lin*, Ehsan Azarnasab*, Faisal Ahmed*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. 2023. 3

[41] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*, 2023. 1, 3

[42] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 4

[43] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023. 1, 3

[44] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[45] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv:2307.08581*, 2023. 3

[46] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023. 3

[47] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 1, 3