From Multimodal LLM to Human-level AI: Modality, Instruction, Reasoning, Efficiency and Beyond

Hao Fei[♣] Yuan Yao[♣] Zhuosheng Zhang[♡] Fuxiao Liu[♠] Ao Zhang[♣] Tat-seng Chua[♣]
[♣]National University of Singapore
[♡]Shanghai Jiao Tong University
[♠]University of Maryland, College Park
haofei37@nus.edu.sg, yaoyuanthu@gmail.com, zhangzs@sjtu.edu.cn,
fl3es@umd.edu, aozhang@u.nus.edu, dcscts@nus.edu.sg

Abstract

Artificial intelligence (AI) encompasses knowledge acquisition and real-world grounding across various modalities. As a multidisciplinary research field, multimodal large language models (MLLMs) have recently garnered growing interest in both academia and industry, showing an unprecedented trend to achieve human-level AI via MLLMs. These large models offer an effective vehicle for understanding, reasoning, and planning by integrating and modeling diverse information modalities, including language, visual, auditory, and sensory data. This tutorial aims to deliver a comprehensive review of cutting-edge research in MLLMs, focusing on four key areas: MLLM architecture design, instructional learning, multimodal reasoning, and the efficiency of MLLMs. We will explore technical advancements, synthesize key challenges, and discuss potential avenues for future research. All the resources and materials are available at https://mllm2024.github.io/COLING2024

Keywords: Large Language Model, Artificial Intelligence, Multimodal Learning, Instruction Tuning, Reasoning, Efficiency Learning

1. Introduction

This year, the whole world has witnessed astonishing advancements in artificial intelligence (AI) to date due to the emergence of large language models (LLMs), such as OpenAI's ChatGPT (OpenAI, 2022b) and GPT-4 (OpenAI, 2022a). LLMs have showcased remarkable capabilities in understanding language, hinting at the not-so-distant arrival of true AGI. Following ChatGPT, a series of open-source LLMs have been published, e.g., Flan-T5 (Chung et al., 2022), Vicuna (Chiang et al., 2023), LLaMA (Touvron et al., 2023a) and Alpaca (Taori et al., 2023), sparking a surge in research revolving around LLMs. The advent of LLMs has also profoundly changed the way tasks are modeled within the NLP community. Human interactions with NLP models have shifted from traditional methods like classification and sequence labeling to a unified 'query-answer' paradigm between user and agent with natural prompt texts (Lester et al., 2021). LLMs have demonstrated promising results in both zero-shot and few-shot settings across various NLP and CV tasks, even with some existing benchmarks being well solved.

However, in reality, we humans inhabit a world where various modalities of information coexist, including visual, auditory, sensory and more, beyond pure language. This realization underscores the necessity of endowing LLMs with multimodal perception and comprehension capabilities to achieve human-level AI, i.e., AGI. This endeavor has given rise to an emerging topic of Multimodal LLMs (MLLMs). MLLMs offer a compelling argument for enhancing the robustness of LLMs by enabling multisensory learning, with each sensory modality complementing the others. Researchers devise additional encoders in front of textual LLMs for receiving inputs in other modalities, leading to the development of MLLMs, such as BLIP-2 (Li et al., 2023), Flamingo (Alayrac et al., 2022a), MiniGPT-4 (Zhu et al., 2023), Video-LLaMA (Zhang et al., 2023c), LLaVA (Liu et al., 2023e), PandaGPT (Su et al., 2023), SpeechGPT (Zhang et al., 2023b) and NExT-GPT (Wu et al., 2023b).

As the manner of interactions with LLMs has been shifted into a more human-centric 'query-answer' style, the learning of LLMs has also been changed. Different from the typical training of deep models, e.g., masked language modeling (Devlin et al., 2019), instruction tuning has been introduced as a major approach for LLMs/MLLMs' tuning (Yin et al., 2023; Su et al., 2023). With sufficient instruction tuning, LLMs/MLLMs are taught to faithfully follow human instructions. Also, it is critical to fully exploit the potential of LLMs/MLLMs for achieving human-level reasoning. Correspondingly, researchers have designed the Chain-of-Thought (CoT) concept (Wei et al., 2022b), which offers a solution enabling LLMs with complex problem-solving abilities on language (Wang et al., 2023; Fei et al., 2023a) or multimodal data (Zhang et al., 2023d; Zhang and Zhang, 2023). Simultaneously, it has been demonstrated that the larger the model

LREC-COLING 2024 Tutorials, pages 1–8
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

sizes and parameters, the more evident the emergence of capabilities in LLMs/MLLMs (Wei et al., 2022a). However, constructing and training extremely large-scale LLMs come at a significant cost, which poses a great challenge for widespread research in this field. Consequently, the efficient development of models becomes a crucial aspect of MLLM's progress.

In this **cutting-edge tutorial**, we aim to offer a comprehensive introduction to techniques for building MLLMs that contribute to achieving stronger, more efficient and more human-level AI. We will delve into recent progress in the realm of MLLMs under four parts, which also are the key components of the topic of MLLMs. **First, multi-modality architecture design**, we elaborate on the cutting-edge approaches to designing architectures that seamlessly integrate multiple modalities, enabling MLLMs to process a variety of sensory inputs effectively. **Second, instruction learning**, we delve into the intricacies of instruction learning, where we discuss the methods and strategies used to train models to follow human instructions under multi-modalities accurately. **Third, multimodal reasoning**, we will present the techniques and methodologies behind multimodal reasoning, which empowers MLLMs to perform intricate reasoning tasks across different modalities with their cognitive capabilities. **Finally, efficiency of MLLMs**, we will give a brief overview of efficient model development, exploring strategies to construct MLLMs that balance performance with computational resources, making them accessible for a wider range of research applications. For each part of the components, we survey the progress and elaborate all the existing techniques on the track, and finally shed light on the future possible directions.

2. Tutorial Outline

This **half-day** (3.5 hours) tutorial presents a systematic overview of recent advancements, trends, resources and also emerging challenges that cover the following topics.

Part 1: Introduction and Overview (10 mins) We begin motivating the topic of MLLMs with the current progress in both academia and industry for achieving the goal of human-level AI. And then we place the emphasis on the key aspects of building successful MLLMs, which bring out the following tutorial content.

Part 2: MLLM Architecture Design (80 mins) We start with the introduction of pre-training language models (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), and then transit to the LLMs of pure languages (OpenAI, 2022a; Touvron et al., 2023b), e.g., ChatGPT. Key techniques of LLMs will be highlighted. Then, we delve into the development of MLLMs based on the success of textual LLMs. We will review the architecture design and training techniques of existing popular MLLMs from two main aspects. (1) First, we will summarize vanilla MLLM architectures that integrate LLMs with different modality information (Alayrac et al., 2022b; Li et al., 2023; Liu et al., 2023e), including multimodal encoding, fusion and generation. (2) Second, we will review the pretraining techniques to learn foundational MLLM capabilities from large-scale multimodal data (Alayrac et al., 2022b; Hu et al., 2023; Radford et al., 2021).

We humans consistently keep engaging in the process of receiving and producing multimodal content every minute and hour, e.g., language, visual, sound, touch and smell. Thus, building MLLMs that only can understand multimodal information is never enough to achieve the goal of human-level AI. In this sub-topic, we further introduce the current progress in developing unified multimodal agents that are able to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, audio, and beyond (Wu et al., 2023a; Shen et al., 2023; Tang et al., 2023; Wu et al., 2023b). We present the existing popular modeling architectures of the any-to-any MLLMs, as well as the discussion in terms of their pros and cons. And finally we shed light on the key points in realizing the more human-like MLLMs, such as the concept of world knowledge modeling, and end-to-end unified agents.

Part 3: Multimodal Instruction Tuning (40 mins) Multimodal instruction tuning typically refers to the process of optimizing instructions or guidance for a system or model that can understand and process multiple types of inputs, such as text, images, audio, etc. Recent open-source instruction-tuned MLLMs including Alayrac et al. (2022a); Zhu et al. (2023); Zhang et al. (2023c); Liu et al. (2023e); Su et al. (2023); Liu et al. (2023b); Zhang et al. (2023b); Wu et al. (2023b); Liu et al. (2023b,d) have shown remarkable performance. In this part, we will delve into how to build instruction-tuned MLLMs step by step. This session is structured as follows. (1) First, we will introduce the construction of visual instruction data and how to improve data quantity and quality. (2) Second, We will engage in the intricate details of the architecture and training strategies of current MLLMs, like MiniGPT4 (Zhu et al., 2023), LLaVA (Liu et al., 2023e) and etc. (3) Third, we will discuss the challenges in this domain, including parameter-efficient training and relieving hallucination issues (Liu et al., 2023c,a).

Part 4: Multimodal Reasoning (40 mins) Imagine trying to study a textbook without any figures, diagrams, or tables. Multimodal reasoning is a rapidly evolving research field that aims to enhance

deep learning models by enabling them to learn from information gathered from various sources and engage in complex reasoning (Hu et al., 2017; Alayrac et al., 2022b; Lu et al., 2022; Yang et al., 2023; Driess et al., 2023). In this section, we will delve into the techniques and methodologies that form the foundation of multimodal reasoning. These techniques empower MLLMs to perform intricate reasoning tasks across different modalities, drawing upon their cognitive abilities. This session is structured as follows. (1) First, we will introduce benchmark datasets and assess the performance of MLLMs on these benchmarks. (2) Second, we will engage in a detailed discussion exploring key research topics, including multimodal chain-of-thought reasoning (Zhang et al., 2023d), multimodal in-context learning (Zhao et al., 2023b), and compositional reasoning (Lu et al., 2023). (3) Third, we will address the challenges faced in this area and discuss future research directions, including multimodal tool learning and multimodal autonomous agents.

Part 5: Efficient MLLM Development (40 mins) MLLM construction (Alayrac et al., 2022b; OpenAI, 2022a) is typically costful, which usually takes thousands of GPU hours and causes severe carbon emissions. In this condition, efficient MLLM development aims at training MLLMs with reduced training cost, while still ensuring excellent multimodal understanding ability. In this section, we will make a systematical review of the techniques that contribute to training efficiency from 3 aspects: (1) First of all, to reduce the training cost, parameter-efficient tuning like LoRA (Hu et al., 2021) is usually employed. We will introduce several parameter-efficient tuning methods (Hu et al., 2021; Dettmers et al., 2023) and corresponding examples. (2) Secondly, using the high-quality training data (Liu et al., 2023e; Li et al., 2023) is essential to boost the training efficiency. We list the widely used databases and make a discussion on their effects. (3) Thirdly, we will introduce how to organize the above mentioned techniques by using different training paradigms. For example, VPG-Trans (Zhang et al., 2023a) propose a two-stage transfer learning framework to realize MLLM construction with around 10% cost. After reviewing existing techniques, we will discuss the challenges and future directions, including how to decide the optimal corpus composition and search for the most efficient training paradigm.

3. Reading List

LLMs and MLLMs. GPT-3 (Brown et al., 2020); GPT-4 (OpenAI, 2022a); Flamingo (Alayrac et al., 2022b); BLIP-2 (Li et al., 2023); LLaVA (Liu et al., 2023e); Visual ChatGPT (Wu et al., 2023a); HuggingGPT (Shen et al., 2023); CoDi (Tang et al., 2023); ImageBind (Girdhar et al., 2023); NExT-GPT (Wu et al., 2023b); AnyMAL (Moon et al., 2023); VisCPM (Hu et al., 2023); Muffin (Yu et al., 2023); Qwen-VL (Bai et al., 2023); KOSMOS-2 (Peng et al., 2023).

Instruction Tuning. MiniGPT4 (Zhu et al., 2023); LLaVA (Liu et al., 2023e); LRV-Instruction (Liu et al., 2023b); Llama-adapter v2: (Gao et al., 2023); SVIT (Zhao et al., 2023a); mplug-owl (Ye et al., 2023).

Reasoning with LLM. Multimodal-CoT (Zhang et al., 2023d); MMICL (Zhao et al., 2023b); Chameleon (Lu et al., 2023); Auto-UI (Zhang and Zhang, 2023).

Efficient Learning. LoRA (Hu et al., 2021), QLoRA (Dettmers et al., 2023), LLaVA (Liu et al., 2023e), LaVIN (Luo et al., 2023), VPGTrans (Zhang et al., 2023a).

4. Presenters

Hao Fei (https://haofei.vip). He is currently a research fellow in the School of Computing, National University of Singapore; and also an associate researcher at Sea AI Lab, Singapore. His research interests cover NLP and multimodal learning, with specific interests in structural learning and LLMs. Over 40 of his research papers have been published at top-tier venues, *e.g.*, ICML, NeurIPS, ACL, ACM MM, AAAI, SIGIR, IJCAI, WWW, EMNLP, TOIS, TNNLS. He won the Paper Award Nomination at ACL 2023. He co-organized the Workshop on Deep Multimodal Learning for Information Retrieval at ACM MM 2023. He has been the co-organizer of top-tier conferences, such as Workshop Chair and Volunteer Chair in EMNLP, WSDM and ACL. He served as Area Chair and Senior Program Committee in relevant multiple conferences, such as EMNLP, WSDM, AAAI, IJCAI and ACL.

Yuan Yao (https://yaoyuanthu.github.io/). He is currently a research fellow in the School of Computing, National University of Singapore. His research interests include MLLMs and information extraction. He has published over 20 papers in top-tier conferences and journals, including ACL, EMNLP, NAACL, COLING, ICCV, ECCV, NeurIPS, AAAI, and Nature Communications. He has served as a PC member for ARR, ACL, EMNLP, NeurIPS, AAAI, WWW, etc.

Zhuosheng Zhang (https://bcmi.sjtu.edu.cn/~zhangzs/). He is currently an Assistant Professor at Shanghai Jiao Tong University, China. His research interests include NLP, LLMs, and multimodal autonomous agents. He has published over 50 papers in top-tier conferences and journals, including TPAMI, ICLR, ACL, AAAI, EMNLP, TNNLS, TASLP, and COLING. He has won 1st place in various language understanding

and reasoning leaderboards, such as HellaSwag, SQuAD2.0, MuTual, RACE, ShARC, and CMRC. He has several tutorials at conferences, including IJCAI 2021 and IJCNLP-AACL 2023.

Fuxiao Liu (https://fuxiaoliu.github.io). He is currently a PhD student in the school of Computer Science, University of Maryland, College Park. His research interests cover multiple vision and language tasks, including image/video captioning, multimodal semantic alignment, fact-checking, document understanding. His recent focus is on building customizable large models that follow humans' intent. His research has been published at top-tier venues, *e.g.*, EMNLP, ICLR, EACL, COLING. He has ever interned multiple companies, including Nvidia, Adobe, Microsoft and Tencent.

Ao Zhang (https://waxnkw.github.io). He is currently a PhD student in the School of Computing, National University of Singapore. His research interests mainly lies on multimodal large language model, multimodal prompt learning and structured scene understanding. He has published several papers on top-tier conferences including ICCV, ECCV, ACL, EMNLP, AAAI, and NeurIPS.

Tat-seng Chua (https://chuatatseng.com). He is the KITHCT Chair Professor with the School of Computing, National University of Singapore, where he was the Acting and Founding Dean of the School from 1998 to 2000. His main research interests include multimedia learning and social media analytics. He is the Co-Director of NExT++, a joint center between NUS and Tsinghua University, to develop technologies for live social media search. He is the 2015 winner of the prestigious ACM SIGMM Technical Achievement Award and has received the best papers (or candidates) over 10 times in top conferences (SIGIR, WWW, MM, etc). He serves as the General Chair of top conferences multiple times (MM 2005, SIGIR 2008, WSDM 2023, etc), and the chief editors of multiple journals (TOIS, TMM, etc). He has given invited keynote talks at multiple top conferences, including the recent one on the topic of large language models.

5. Other Information

Type of Tutorial: Cutting-edge.

Past Tutorials: To our knowledge, there is no prior tutorial for delivering comprehensive instruction on the topic of multimodal LLMs.

Target Audience: Our tutorial is targeted at members of a broad range of relevant communities, e.g., NLP, CV and broad AI, who have interests in building LLMs and applying LLMs to achieve stronger task performances. This includes researchers, students of both academia and industry, as well as practitioners wishing to make use of LLMs in their learning pipelines. We expect that participants are comfortable with the basic foundations of both NLP and multimodal learning tasks, as well as the basic knowledge of standard generative models e.g., transformers. While we do not require any readings, we recommend reviewing the works cited in this proposal, especially the reading list.

Prerequisites: Following knowledge is assumed:

- Machine Learning: basic probability theory, supervised learning, transformer models
- NLP: Familiarity with LLMs; prompt tuning technique, generative NLP, etc.
- Multimodal Learning: Familiarity with multimodal modeling, e.g., visual, video, audio; diffusion models, etc.

Estimated Participant Number: 200.

Breadth: We estimate that approximately 30% of the tutorial will center around work done by the presenters. This tutorial categorizes the goal of developing successful MLLMs into several sub-topics, and each of the sub-topics includes a significant amount of other researchers' works.

Open Access: We make all teaching material available online, and we agree to allow the publication of slides and video recordings in the LREC-COLING 2024.

Diversity Considerations: The content and methods in this tutorial broadly cover the key common knowledge from NLP, CV and machine learning fields. Thus, this tutorial will facilitate a wide range of communities in diverse topics and domains. The speakers are from diversified academic institutions with different backgrounds and regions, e.g., including both professors, research fellows and Ph.D. students, and from Sinagpore, China and USA. We will reach out to academic communities to encourage them to attend our tutorial for the participation of diverse audiences.

6. Ethics Statement

Our tutorial is committed to promoting the research and responsible AI development. All the materials cited, occurred and presented in this tutorial strictly follow the corresponding regulations and licenses. We emphasize the importance of respecting user privacy, ensuring fairness in LLM systems, and advocating addressing potential biases across modalities. We encourage participants to consider the societal impact of their work and prioritize transparency, accountability, and inclusivity in their research. Together, we aim to advance multimodal AI technologies while upholding the highest ethical standards.

7. Bibliographical References

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

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2023. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. *arXiv preprint arXiv:2311.18651*.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023a. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1171–1182.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023b. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5980–5994.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. 2023c. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 804–813.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pages 19730–19742.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023c. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023d. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning. *CoRR*, abs/2304.08485.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023f. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.

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*, 35:2507–2521.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

OpenAI. 2022a. Gpt-4 technical report.

OpenAI. 2022b. Introducing chatgpt.

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

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable

visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2609–2634.

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. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

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

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. 2023. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687.

Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. 2023. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *CoRR*, abs/2305.11000.

Hang Zhang, Xin Li, and Lidong Bing. 2023c. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858.

Zhuosheng Zhang and Aston Zhang. 2023. You only look at screens: Multimodal chain-of-action agents. *ArXiv preprint*, abs/2309.11436.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023d. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Bo Zhao, Boya Wu, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023b. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Li Zheng, Hao Fei, Fei Li, Bobo Li, Lizi Liao, Donghong Ji, and Chong Teng. 2023. Reverse multi-choice dialogue commonsense inference with graph-of-thought. *arXiv preprint arXiv:2312.15291*.

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